VARIABLE-LENGTH COMPUTERIZED ADAPTIVE TESTING: ADAPTATION OF THE *A*-STRATIFIED STRATEGY IN ITEM SELECTION WITH CONTENT BALANCING

BY

YAN HUO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2009

Urbana, Illinois

Doctoral Committee:

Professor Hua-Hua Chang, Chair
Professor David V. Budescu, Director of Research
Professor Lawrence J. Hubert
Professor Carolyn J. Anderson
Professor Jeffrey A. Douglas

Abstract

Variable-length computerized adaptive testing (CAT) can provide examinees with tailored test lengths. With the fixed standard error of measurement (*SEM*) termination rule, variable-length CAT can achieve predetermined measurement precision by using relatively shorter tests compared to fixed-length CAT.

To explore the application of variable-length CAT, this dissertation proposes four variable-length item selection methods adapted from the *a*-stratified strategy (Chang & Ying, 1999). These methods are named 1) the circularly increasing *a*-stratified method (STR-Ca), 2) the circularly decreasing *a*-stratified method (STR-Cd), 3) the random *a*-stratified method (STR-R), and 4) the two-stage *a*-stratified variable-length method (STR+R). The general strategy of these four methods allows test items to be selected in a mixed-strata ordering fashion from all strata partitioned by different levels of the discrimination parameter. This flexibility can overcome the potential problem of unbalanced item usage across different strata caused by previous attempts of applying the original *a*-stratified method into variable-length CAT.

Study 1 examines the STR-Ca, the STR-Cd, and the STR-R methods in fixed-length CAT situations and the results show that their performance is comparable to that of the original *a*-stratified method in the fixed-length simulations in terms of various criterion measures such as Bias, *MSE*, efficiency, and item exposure rates. Study 2 explores these four item selection methods under the variable-length situations and the results indicate that these four methods can achieve good ability estimation while maintaining balanced item usage in the variable-length CAT simulations. To extend the implementation of these four variable-length item

selection methods into a more realistic testing situation with content balancing constraints, Study 3 proposes two two-phase content balancing control methods, the variable-length modified multinomial model (MMM) method and the content weighted item selection index method. They can be naturally incorporated with these four adapted a-stratified methods to realize variable-length CAT with content control. Lastly, the intent of Study 4 is to explore decision making tools regarding choices among several variable-length CAT designs. Two quantitative indices, the cost-effective ratio and the variable-fixed-fitness index, are developed and their applications are demonstrated with some hypothetical examples.

Together, these study findings will advance the research and understanding of variable-length CAT, and will facilitate the application and adoption of variable-length CAT in real world testing.

Acknowledgements

Working on my dissertation is a long journey where I have to walk mostly on my own. However, the accomplishment would be impossible without many helps and supports from other people.

Many thanks to my dissertation advisor Dr. Hua-Hua Chang for inspiring me to explore my dissertation study in the area of computerized adaptive testing. His extraordinary expertise knowledge in theoretical and technical aspects always guided my research on the right track. I would like to express my sincere gratitude to the director of research of my dissertation, Dr. David V. Budescu. His insightful comments, constructive criticisms and revisions on writing are the invaluable feedbacks and encouragement for me to enrich my dissertation. My academic advisor, Dr. Carolyn J. Anderson has always been there for listening and giving me advice. I am very thankful for her careful reading and grammar advice on my dissertation manuscripts. I sincerely thank Dr. Lawrence J. Hubert, who has never been hesitated to offer help and advice, and to guide me to use the APA style and improve my writing skills. I also thank Dr. Jeffrey A. Douglas for being my committee member and his astute suggestions to improve the quality of my work.

I am thankful to Dr. Carol Nickerson, the coordinator at the quantitative division. She made my remote study much more smoothly. I am also grateful to Lori Hendricks. She was always the first person whom I would bother if I had any concerns during my graduate study.

I would like to thank many friends I have known these years. Their friendships added bright colors to my graduate life in Urbana-Champaign.

Most importantly, none of this would be completed and meaningful without the love, support and understanding from my dearest family. I would like to express my heartfelt gratitude to my parents, my husband and my baby Alan. Specially, this dissertation is dedicated to my father and my mother.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

This introductory chapter gives an overview of the general ideas of computerized adaptive testing (CAT), including its background and development. Three aspects of CAT most relevant to the research questions in this dissertation are the stopping rule (i.e., fixed versus variable-length design), item selection methods, and content balancing. Subsequently, several research questions are stated.

1.1 Overview of CAT

Computerized adaptive testing (CAT), based on Item Response Theory (IRT), offers tailored tests, where items presented to each examinee are sequentially selected from a large item bank according to the current estimate of the examinee's ability calibrated by an IRT model based on preceding responses. Examinees with high ability can avoid being administered too many relatively easy items, and less proficient examinees do not have to encounter unsuitably challenging items. Compared to the traditional IRT-based paper-and-pencil (P&P) tests, that provide examinees of different ability levels with identical testing versions, CAT effectively improves the precision and efficiency of the ability estimation (Chang, 2004; Weiss, 1982). In addition, there are some appealing non-statistical features of CAT. For example, CAT can provide examinees with more flexible test time, and can produce immediate test scores for reporting to test takers (Meijer & Nering, 1999).

In essence, adaptive testing is an interactive testing procedure that allows appropriate items to be administered to each examinee according to ability/trait level. Such an adaptive assessment is not the original invention of modern CAT. The pioneer who first applied the

adaptive testing philosophy into real test situations was Alfred Binet, a French psychologist and the inventor of the first intelligence test, the Binet IQ test (Binet & Simon, 1905). With no computer technology assistance, he created an adaptive intelligence testing procedure that included basic adaptive ingredients consistent with the view of modern adaptive testing (Weiss, 2004). For example, a large item bank was constructed beforehand. The test starting level could be tailored for different examinees because the tester may "guess" examinees' intelligence level according to certain cues, such as the child examinee's chronological age. Based on each child's performance on a set of intelligence questions, the next set of items was selected. A specific termination rule involving "ceiling" and "basal" levels was also taken into consideration to stop the test appropriately.

Building upon Binet's intelligence adaptive test, a number of large-scale theoretical and applied studies on adaptive testing as well as its relevant issues were conducted in the U.S (e.g., Lord, 1970, Weiss & Betz, 1973) from the late 1960s. Meanwhile, some initial studies about non-IRT-based adaptive testing were explored, such as Robbin-Monro procedure, fixed step size, flexilevel test, Bayesian procedures, and stratified adaptive test (see Rudner,1978). Lord (1970) proposed the maximum information method, which can be used in IRT-based adaptive testing. This approach selects the item with the maximum Fisher item information evaluated at the current ability estimate. Weiss (1982) pointed out that IRT-based CAT can produce more desirable results than non-IRT adaptive tests because IRT approach can provide a meaningful expression of ability scores on a same metric.

In the last three decades, with the widespread accessibility of computers, CAT (referring to the IRT-based CAT for the remaining part) has been extensively applied in large-scale measurement and testing situations. The Armed Services Vocational Aptitude Battery (ASVAB) was the first well-known CAT application. Other operational CAT applications are the Graduate Record Examination (GRE), the Graduate Management Admission Test (GMAT), the American Institute of Certified Public Accountants (AICPA), and the National Council Licensure Examinations (NCLEX). Psychologists also introduced the CAT approach into the study of personality and attitude assessments (for example, see Waller &

Reise, 1989, Dodd, De Ayala, & Koch, 1995, Craig & Harvey, 2004). As CAT has become a popular testing tool in recent years, some issues regarding the large-scale applications of CAT have unexpectedly emerged. To name several influential events, in the year 2000, thousands of GRE CAT takers had unreliable GRE scores and were offered a retake without additional charge by the Educational Testing Service (ETS). Later in 2002, the GRE CAT was replaced by the P&P version in some countries and areas due to the test security problem. Such setbacks for CAT applications stimulated researchers to address these issues and seek solutions (for example, Chang & Ying, 2002).

1.2 Item selection methods

The essence of CAT is the adaptive nature of sequentially selecting items most appropriate to measuring an examinees' ability. Thus, one of the most essential components to design a CAT system is the item selection procedure.

1.2.1 The Maximum-information method

One of the popular IRT models used in CAT application is the three-parameter logistic ($3PL$) model:

$$P(Y_j = 1|\theta) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j}(\theta - b_j)} \tag{1.1}$$

where

θ is the ability/trait level of each examinee;

$P(Y_j = 1|\theta))$ is the item response function for item j; that is, the probability of answering item j correctly given θ;

a_j is the j^{th} item discrimination (slope) parameter;

b_j is the j^{th} item difficulty (location) parameter;

c_j is the j^{th} guessing (asymptote) parameter;

The maximum information (MI) method (Lord, 1970) is an information-based item selection procedure to select items with maximum Fisher information. Under a simple case of

$c = 0$, the $3PL$ IRT model is equivalent to $2PL$ model, and Fisher information is expressed as

$$I_j(\theta) = \frac{a_j^2 e^{a_j(\theta - b_j)}}{[1 + e^{a_j(\theta - b_j)}]^2}. \tag{1.2}$$

In this case, maximizing Fisher information implies matching the value of the item difficulty parameter (b) with the latent trait level (θ) of each examinee. Because the latent trait level is unknown, the MI method approximates this ideal process by matching b with the provisional estimated latent trait level ($\hat{\theta}$). The reciprocal of the Fisher information is the asymptotic sample variance of $\hat{\theta}$ implying that such maximization minimizes the variance of the ability estimate ($\hat{\theta}$) and makes $\hat{\theta}$ the most efficient estimator. The same results are valid for the $3PL$ model under some regularity conditions (Chang, 2004). For this reason, the MI method has been the most common item selection method over the last several decades.

It is possible to show with formula 1.2 that information reaches maximum value $a^2/4$ at $b = \theta$, thus, items with high a values have high information. The MI method tends to select items with b close to θ and an a as large as possible, leading to skewed item usage within the item pool. Thus, some "good" items with high discrimination are frequently selected while other "bad" items with lower discrimination are unfortunately neglected. The item overexposure/underexposure phenomena may cause test security problems and item pool usage inefficiency. In practice, an item pool only contains several hundred items. Overexposing popular items increases the test security risk, especially in high-stakes test settings, where popular items that are frequently administered in a relatively short period may "leak"(Mills & Stocking, 1996). Item underexposure is economically inefficient to test makers as the construction of an item bank involves substantial financial investments.

1.2.2 Item exposure control methods

The key to solving the item over/under exposure issue caused by using the MI method is item exposure control. A number of item selection methods focusing on controlling item exposure have been proposed in recent years. Georgiadou and Triantafillou (2007) reviewed thirty-

one different item selection strategies handling item exposure control from 1983 to 2005, and classified them into five main categories: (1) randomization, (2) conditional selection, (3) stratified strategies, (4) combined strategies, and (5) multiple stage adaptive test design.

The general principle of the randomization strategies is the random selection of an item from a group of most informative items. Most randomization strategies add a random component to the maximum information item selection procedure. For example, the 5-4-3-2-1 strategy proposed by McBride and Martin (1983) introduces randomization processes into the initial stage of item administration. It makes a random selection of the first item from a group of five most informative items and randomly selects the second item from a group of four most informative items and so on untill a random selection of the forth item from two most informative items. Starting from the fifth item, items are selected solely according to the maximum information criterion. The randomization process at the initial four items can reduce the over-exposure rates for the most informative items at the early testing stage.

The mainstream conditional strategies employ a probabilistic exposure control parameter to control the exposure rate for each item. One of the most common conditional selection methods is the Sympson-Hetter strategy (Sympson & Hetter, 1985) that differentiates the item selection and the item administration procedures. The Sympson-Hetter (S-H) method assumes

$$P(A_j) = P(A_j/S_j)P(S_j), \tag{1.3}$$

where

$P(S_j)$ is the probability of selecting item j;

$P(A_j)$ is the probability of administering item j;

$P(A_j/S_j)$ is the control parameter, the probability of administering an item given that it is selected.

$P(A_j)$ can be effectively controlled by manipulating $P(A_j/S_j)$. Suppose the exposure rate is r_j for item j. The exposure rate is defined as the ratio of the number of times the item is

administered to the total number of examinees. If $P(A_j/S_j) \leq r_j/P(S_j)$, then $P(A_j) \leq r_j$. Item j selected according to the maximum information criterion does not guarantee it will be administered. Whether item j is administered depends on the conditional probability $P(A_j/Sj)$ being less than some criterion: $r_j/P(S_j)$.

The control parameters have to be estimated by computing algorithms before operational implementations. The values of the control parameters are not population invariant and will change as item drift occurs. Therefore, generating and maintaining appropriate control parameters are computational demanding and time consuming tasks. In addition, the S-H method can only effectively control overexposure of those items whose exposure rates exceed the preset r, but it has no power to increase the underexposure rates.

The stratified, combined strategies, and multiple stage adaptive test design are relatively new developments in the item exposure control field (Georgiadou & Triantafillou, 2007). The stratification approach divides the entire item pool into several strata according to the ascending values of a. The selection algorithm is performed in each stratum consecutively during the course of the testing, and thus, the item exposure rates of items can be optimally equalized. The original method, a-stratified multistage method (Chang & Ying, 1999) and its derivations will be explained in the next section.

Different item control methods can be combined to solve item exposure problems in a better way than using one strategy alone. A number of hybrid methods were proposed based on some conditional and stratification approaches. For example, Sympson-Hetter control can be combined with the a-stratified multistage method (Yi, 2002; Leung, Chang & Hau, 2002).

A standard multiple stage adaptive test provides examinees with an adaptive test at a level of subtests or testlets. This kind of test usually incorporates some P&P test features, such as parallel forms and review by test specialists, to overcome some CAT problems such as exposure control (Armstrong & Little, 2003; Armstrong & Edmonds, 2004).

1.2.3 The a-Stratified multistage method

Chang and Ying (1999) proposed the a-stratified multistage method aiming to equalize exposure rates. In this method all items in the item pool are stratified with respect to the levels of a (the discrimination parameter). The low-a items are administered at the earlier stages of the test and the items with high discrimination are saved for the latter stages. Their simulation results showed that this approach can effectively control the overexposure rates for the high-a items while substantially increasing the administration rates for the low-a items.

A simple a-stratified multistage procedure used in a fixed-length CAT (test length is denoted as L) design is described as follows.

Step 1: Sort the values of a in an ascending order.

Step 2: Partition the item pool into S strata based on the sorted values of a.

Step 3: Partition the test into S stages.

Step 4: In the s^{th} stage ($s = 1, \ldots, S$), select L/S items from the corresponding s^{th} stratum sequentially based on a chosen item selection criterion, such as selecting an item whose b is closest to $\hat{\theta}$.

Step 5: Repeat step 4 utill the test is finished.

Unlike the S-H method, the a-stratified method is a design-like approach. It is relatively easy to implement in operational testing situations. Furthermore the a-stratified method is advantageous over the S-H method on enhancing underexposure rates. It can effectively increase the exposure rates of the items with low discrimination and control the overexposure rates for highly discriminating items.

Basically, the a-stratified method is an a ascending item selection method in contrast to other methods, such as the MI method that tends to select items with larger a items at the earlier stage of testing (Chang & Ying, 2008, Hau & Chang, 2001). There are solid theoretical grounds to support the use of the ascending approach. It is clear that items with high a and b close to $\hat{\theta}$ provide the most information in the $2PL$ and $3PL$ models

(Hambleton & Swaminathan, 1985), but how to allocate items with different a values during the course of testing is an important issue. Equation 1.2 shows that information reaches its maximum value at $a^2/4$. The true θ is never known in practice, but a close approximation is to use $\hat{\theta}$ instead. If $\hat{\theta}$ is far from θ (this is a common case at the earlier stage of testing when only a small number of items are available to estimate θ), the information value can be much smaller than the expected $a^2/4$, and hence, the maximum information criterion can be inefficient. The cause of this inefficiency problem is that the Fisher information is the local information requiring extensive knowledge about the location of θ (Chang & Ying, 1996). As a result, the usefulness of an item with high discrimination cannot be fully optimized at the beginning of the test when $\hat{\theta}$ is usually not very accurate (see Chang & Ying, 1999, Chang & Ying, 1996). Thus, items with high discrimination are used more efficiently when saved for latter testing stages, where $\hat{\theta}$ can be estimated based on responses to more items, and therefore, more likely to be close to θ.

The preceding description of the 5-step a-stratified procedure is an elementary version proposed by Chang and Ying (1999). This prototype has been flexibly modified in many ways to solve more complex CAT questions. In response to Stocking's (1998) concerns regarding the correlation between the item difficulty and item discrimination, Chang et al. (2001) developed the a-stratified multistage with b blocking method. This approach is appropriate for the tests that contain items with fairly or strongly correlated a's and b's. This refined approach can balance the distribution of b among all strata so that the effect of b that might interact with the stratification can be eliminated.

The a-stratified multistage method can also be combined with other item selection methods. When the ratio of item pool size to test length is too small, the a-stratified method does not control overexposure effectively. Leung et al. (2002) combined the a-stratified multistage method and the S-H algorithm to enhance exposure control effectiveness. Leung et al. (2000a) integrated the weighted deviation model (Stocking & Swanson, 1993) into the a-stratified method and the a-stratification with blocking method, respectively.

Compared with the MI method, the a-stratified design sacrifices test efficiency to some

degree to control item exposure rates. This is also true for the S-H method. Deng and Chang (2001) modified the a-stratified design into a stratification procedure with unequal item exposure across strata. Specifically, this method allows fewer items to be selected from low discrimination strata with more items selected from high discrimination strata.

To summarize, the a-stratified multistage method is a sampling approach that partitions the item pool according to the ascending values of the discrimination parameter. It aims to equalize item exposure rates (for both over- and under- exposed items) while maintaining test efficiency. It is easy to implement and flexible enough to be modified for different circumstances. Also, the a-stratified multistage method can incorporate some non-psychometric constraints, such as content balancing. This refined method will be discussed in part 4.

1.3 Fixed-length versus variable-length CAT

Traditional P&P tests have fixed length that is determined in advance for all examinees who take the same test. In contrast, due to its adaptive nature, CAT can offer either fixed-length tests or variable-length tests. Thus, the choice of stopping rules is one of essential components of CAT.

A fixed-length computerized adaptive test is terminated when a predetermined test length is reached regardless of concerns for the measurement precision or other matters. Stopping a variable-length test does not depend on the specific number of items administered. There are two common variable-length stopping rules. One stopping rule controls the precision of ability estimates. A straightforward application is the fixed standard error of measurement (*SEM*) stopping rule. It requires that a certain prespecified level of measurement precision is achieved. The length of test usually varies for different examinees because more or less items may need to be administered to different test takers to reach certain prespecified measurement precision. Letting σ be the standard error of the ability estimate, and ϵ the predetermined cut-off value, this fixed *SEM* rule can be expressed as follows:

$$\sigma_e \leq \epsilon. \tag{1.4}$$

Based on the Fisher test information function being the reciprocal of the asymptotic sample variance of the ability estimator,

$$I_n(\theta_n) = \frac{1}{\mathrm{Var}(\hat{\theta}_n)} \text{ as } n \to \infty. \tag{1.5}$$

Therefore, the Fisher test information can be directly used as the stopping criterion (e.g., see Wen et al. 2002). The corresponding stopping rule is expressed as

$$I_n(\theta_n) \geq \frac{1}{\epsilon^2}. \tag{1.6}$$

A variation of this latter information stopping rule was proposed recently (Chang & Ying, 2004; Chang & Martinsek, 1992; Grabovsky & Chang, 2001). Besides maintaining the uniform measurement error, this stopping rule constructs a confidence interval for the true value of θ. Let d be the predetermined width of the confidence interval and $z_{\alpha/2}$ the $(1 - \alpha/2^{th})$ quantile of the standard normal distribution. The stopping rule is written as

$$I_n(\theta_n) \geq \left(\frac{z_{\alpha/2}}{d}\right)^2. \tag{1.7}$$

Both stopping formulas (1.6) and (1.7) control the level of measurement precision and realize the uniform measurement precision for different examinees in one test. The latter provides a meaningful interpretation that constructs confidence intervals for the true value of ability.

Another alternative is the confidence interval stopping rule. This procedure is specifically applied in pass/fail decisions, such as classification assessment or adaptive mastery testing (Waller & Reise, 1989). The test continues until a required confidence interval of the current ability estimate does not contain (is either above or below) the pre-established cut-off score (Bergstrom & Lunz, 1992; Kingsbury & Weiss, 1983; Thompson, 2007).

Thissen and Mislevy (2000) listed the advantages of variable-length CAT with the fixed *SEM* over the fixed-length CAT: (1) They ensure score estimates that conform to the "equal measurement error variance" assumption of the traditional test theory, and (2) They allow subsequent statistical analyses involving measurement error to be conveniently handled.

Variable-length CAT has not been as widely applied as fixed-length CAT in educational testing and measurement. Test takers may subjectively sense unfairness in selection and decisions if different people receive various test lengths. Tonidandel and Quiones (2002) reported that extremely short tests (using non fixed-length stopping rules) may affect examinees' fairness perceptions. Second, it is more difficult and complex to incorporate other statistical or non-statistical constraints into the variable-length CAT designs. For instance, an ideal fixed-length CAT conventionally requires that the item pool size to be 12 times the length of CAT (Stocking, 1994). An ideal item pool size for a variable-length design with fixed *SEM* cannot be summarized as such a straightforward rule-of-thumb, because it needs to consider many test administration factors, such as the targeted probability of correct response, desired precision of ability estimate and item exposure control as well as the interactions between those factors and examinee characteristics (Iramaneerat & Stahl, 2007). Currently, for some high-stakes tests such as K-12 assessment, people prefer to use conventional fixed-length adaptive tests (Way, 2005).

Despite the differences between fixed-length and variable-length adaptive tests, adaptive tests can be designed as a hybrid combining these two stopping types (Wood & Zhu, 2006). As an example, a test is terminated according to a specific variable-length stopping rule with the constraints that the administered number of items should be above the minimum test length and not beyond the maximum test length.

1.4 Content balancing methods

A successful CAT application not only includes basic psychometric ingredients, such as item selection methods, stopping criteria, item calibration and so on, but it should also take many

non-psychometric and practical constraints into consideration. Some typical constraints are content balancing, item type and answer key balancing. Content balancing refers to the specification that a test should select an appropriate proportion of items from each content area according to the test blueprint. For example, one particular mathematical test may be constrained to have 30% addition items, 30% subtraction items as well as 20% items for each multiplication and division type. Content control has never been a problem in the traditional P&P form because all examinees receive one uniform test version and items with appropriate content can be balanced beforehand by test construction specialists. CAT offers dynamic and customized tests. Without content control, different examinees may confront different tests composed of items from totally different content areas. Such an outcome is unacceptable for both test makers and examinees. It may cause legal challenges and discourage acceptance of CAT. The initial idea of controlling content balancing was addressed by Green, Bock, Humphries, Linn, and Reckase (1984, see p350) and other researchers (e.g., Thissen & Mislevy, 1990, Wainer, et al., 2000). Item selection methods developed so far for CAT do not automatically incorporate content balancing. Thus, a number of methods have been proposed to achieve content balancing.

Wainer and Kiely (1987) developed an adaptive testlet model, using multi-items "testlets" as the unit for test development and administration. One of its applications can handle content area balancing. Such a test contains a number of subsets measuring different content areas respectively. Kingsbury and Zara (1989, 1991) proposed the constrained CAT (CCAT) method, restricting the item selection algorithm to the targeted content area. Specifically, this method imposes pre-check to locate the content area farthest below its administration percentage, and then applies some item selection algorithm, such as seeking the highest information item for this content area. CCAT can guarantee content control but sacrifices test efficiency to some degree.

One potential problem of the CCAT method is that the sequence of content areas to be selected is predictable, and thus, generates unfavorable order effects. Two other methods were proposed to eliminate order effects. The modified multinomial model (MMM) method

developed by Chen et al. (1999) introduces a random mechanism to eliminate the predictable content sequence. This method produces a cumulative distribution based on the target proportions of the content areas. Next, a random number from a uniform distribution $U(0, 1)$ is generated as a pointer corresponding to the content area where the next optimal item should be selected. Once a content area has fulfilled its target percentage, the multinomial distribution is updated by adjusting the remaining available content areas. The other remedy is the modified CCAT method (MCCAT, Leung, Chang, & Hau, 2000 b) that relaxes the strictness of selecting the content area that is farthest below its target percentage. MCCAT allows optimal items to be selected from any content area which is unfulfilled with available quota. This flexibility helps eliminate the undesirable order effect.

Leung, Chang, and Hau (2003a) compared the performance of CCAT, MMM and MCCAT under the MI with the SH control item selection algorithm by manipulating the two factors of the test length (3 levels) and the target exposure control rate (two levels). Their results showed that all three methods produced comparable estimation accuracy and precision while the MMM method did a better job in overexposure control than the other two methods.

From a different perspective, Yi and Chang (2003) tackled the content balancing issue by refining the a-stratified with blocking method. In addition to partitioning the entire item pool based on a and b, content balancing is added as the third factor. As a result, the item pool is partitioned into strata with different levels of discrimination and each layer contains similar content coverage and similar b distribution. Their simulation results indicate that this modified a-stratified method with content blocking (STR_C) effectively balanced content compared with the original a-stratified multistage method (STR_A), a-stratified multistage with blocking (STR_B) and Fisher information with S-H method.

A comprehensive study involving the above three content balancing approaches and three stratification methods was conducted by Leung et al. (2003b). They evaluated the performance of CCAT, MCCAT, and MMM under three stratification conditions (STR_A, STR_B, STR_C). They recommended that implementing the MMM method under the STR_C stratification condition might be an optimally integrated item selection method in terms of item

overlap control, item pool utilization and content balancing.

The content balancing approaches reviewed so far were developed based on fixed content specification, i.e., the number (or percentage) of items from each content area is fixed. If the number (or percentage) of items from each content area is allowed to vary between a lower bound and an upper bound, we deal with flexible content balancing. The flexible-content balancing in fixed-length CAT can be expressed by the following two formulas:

$$l_k \leq n_k \leq u_k, \tag{1.8}$$

$$\sum_{k=1}^{K} n_k = L, \tag{1.9}$$

where n_k is the number of items that will be actually extracted from each content area k, and l_k and u_k are the lower and upper limits for each content area k. K is the total number of content areas.

On the basis of the MMM and MCCAT approaches, Cheng and Chang (2007) developed four new methods to handle flexible content balancing in the fixed-length CAT situation. Modifying the content specification according to the above formula for the MMM and the MCCAT can realize flexible content balancing. Besides these two modifications, a two-phase strategy is developed, where, two stages of content specifications are formed to meet the requirement of the lower bound an upper bound respectively. This strategy can be applied in the MMM method and the hybrid of MMM and MCCAT (Cheng & Chang, 2007).

Chapter 2

Research Questions

2.1 Research questions

An ideal adaptive test can provide each examinee with a tailored test of a certain test length that may be different from others'. Variable-length testing is a unique advantage of CAT over traditional P&P tests. Motivated by the fact that variable-length CAT has received less attention in both research and application, the central interest of this dissertation is to explore variable-length CAT designs with *SEM* in the framework of the *a*-stratified item selection strategy.

To date there have been relatively few studies on how to modify the *a*-stratified multistage method into variable-length CATs. Wen et al. (2002) first adapted the *a*-stratified multistage (STR) procedure to the variable-length CAT format. Their method partitions the targeted test information into several segments and assigns each of them into each stratum. Once the preset partitioned information is achieved in its stratum, the test moves to the next stage. This continues until the entire targeted test information is reached. In their study, test information could be divided evenly across the strata, or be divided unevenly so less information is allocated in the earlier stages and more information in the latter stages, or vice versa. Uniform information division corresponds to the fixed-length STR method so the number of items extracted from each stratum are equal and without predetermined information restriction. Decreasing or increasing information divisions allow more/less items selected from the earlier test stages and less/more items from the latter test stages; correspondingly, their fixed-length counterparts are the modified STR methods with the unequal usage of strata.

One potential drawback of this adaptation is the uneven item exposure control on different strata for the uniform/increasing/decreasing information division. Partitioning test information evenly can produce fixed and uniform stratum information achievement, but it does not guarantee the numbers of items selected from each stratum to be roughly equal. Furthermore, the decreasing information method largely amplifies this discrepancy. Although the increasing information method reduces the amount of uneven item exposure to some extent in the variable-length design, it cannot entirely eliminate this problem.

The original STR method selects items from strata in a stringent ascending order fashion, where selection can only process to the subsequent stratum until the selection from the preceding stratum is fulfilled. The number of items to be selected from each stratum can be easily predetermined in fixed-length tests but it is not possible to decide for variable-length tests. To facilitate the application of the *a*-stratified method in variable-length CAT with *SEM*, this dissertation proposes four new variations of the *a*-stratified multistage procedure that can be directly applied in variable-length CAT in the fixed *SEM* settings to minimize unbalanced item exposure without partitioning the preset test information.

Three adapted *a*-stratified methods are proposed in Study 1 to demonstrate a flexible multistage selection strategy allowing items to be selected from various strata in a mixed-strata ordering fashion. These three new methods are named (1) the circularly increasing *a*-stratified multistage (STR-Ca) method, (2) the circularly decreasing *a*-stratified multistage (STR-Cd) method, and (3) the randomized *a* -stratified multistage (STR-R) method. The STR-Ca method can be treated as a series of miniature *a*-stratification procedures. STR-Cd is a series of miniature *a*-stratification procedures in the reversed order. STR-R incorporates the randomization component into the *a*-stratified multistage method. Study 1 compares the performances of these three methods to the original STR method, the STR method with unequal exposed strata, the MI method, and the randomized item selection method in the fixed test length scenario.

These three adapted methods and a two-stage *a*-stratified method are examined in the variable-length testing situation in Study 2. The two-stage *a*-stratified (STR+R) method

proposed in Study 2 is a combination of the a-stratification strategy and the randomization process in the STR-R method. It can be applied to variable-length tests with a minimum test length restriction. Study 2 compares their performance to the STR method with unequal exposed strata, the MI method, and the randomization item selection method in the variable-length simulations on the basis of the ability estimation, efficiency, and item exposure control.

Also motivated by the fact that little research on content balancing control has been conducted for variable-length CATs, Study 3 incorporated content balancing constraints for the variable-length CATs. The fixed content balancing specification commonly used for the fixed-length CATs is no longer valid in the variable-length situations. As an alternative, flexible content balancing is appropriate to handle the variable-length feature. Generally, realizing variable-length content balancing control requires the fulfillment of the following three inequalities:

$$l_k \leq n_k \leq u_k, \tag{2.1}$$

$$L \leq \sum_{k=1}^{K} n_k \leq U, \tag{2.2}$$

$$\frac{1}{I} \leq \epsilon^2, \tag{2.3}$$

where n_k is the number of items that will be actually extracted from each content area k; l_k and u_k are the lower and upper limits for each content area k; L and U are the minimum and maximum test length for the variable-length CAT; K is the total number of content areas; and I is the predetermined test information.

Two content balancing control methods proposed in study 3 are the variable-length MMM method and the content weighted item selection index method. These two methods are combined with the STR-Ca, the STR-Cd, the STR-R, and the STR+R item selection method to realize variable-length CAT with content constraints.

The final study deviates from the three previous studies on the proposed stratification methods, Study 4 is more focus on decision making aspects of variable-length designs. A particular fixed-length test may correspond to more than one variable-length alternative.

To deal with such choices, two indices are proposed in Study 4. The cost-effective ratio is derived from the cost-effectiveness analysis and the variable-fixed-fitness index is a composite measure that aggregates several essential performance evaluation criteria into a single quantity to measure the overall performance of the variable-length CAT designs.

2.2 Performance criteria

All three studies employ the performance evaluation criteria commonly used in the previous studies (Chang & Ying, 1999; Wen et al., 2002) to evaluate the performance of various fixed-length or variable-length item selection methods.

The formula to measure the test efficiency is:

$$\text{Efficiency} = \frac{\sum_{i=1}^{m} \text{inf}_i}{\sum_{i=1}^{m} L_i}, \tag{2.4}$$

where m refers to the total number of examinees, L_i is the test length of the i^{th} examinee, and inf_i is the test information of the i^{th} examinee.

The accuracy measurement of ability estimation uses the Bias and Mean Square Error (*MSE*):

$$\text{Bias} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i), \tag{2.5}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \theta_i)^2, \tag{2.6}$$

where θ_i and $\hat{\theta}_i$ are the true and estimated ability of the i^{th} examinee.

The index of item exposure rates in fixed length tests is:

$$\chi^2 = \frac{\sum_{j=1}^{N}(\text{er}_j - L/N)^2}{L/N}, \tag{2.7}$$

where N is the number of total items in the item pool, and er_j is the j^{th} item exposure rate and equals c/m, where c is the number of times the j^{th} item used.

The index of item exposure rates in variable-length tests is:

$$\chi^2 = \frac{\sum_{j=1}^{N}(er_j - \sum_{j=1}^{N} er_j/N)^2}{\sum_{j=1}^{N} er_j/N}. \tag{2.8}$$

Another commonly used index to measure item exposure control is the test overlap rate, defined as the expected number of common items being exposed to any two randomly selected examinees, divided by the test length. Let C_o be the number of common items for any two examinees. There are C_m^2 pairs of M examinees, so the overlap rate can be calculated as:

$$\frac{(\sum_{m} C_o)/C_m^2}{(\sum_{i=1} L_i)/m}. \tag{2.9}$$

Chapter 3

Study 1

Study 1 proposes three stratification methods developed from the a-stratified multistage (STR) method. Specifically, Study 1 evaluates the performance of two circular a-stratified multistage (STR-Ca and STR-Cd) methods and the randomized a-stratified multistage (STR-R) method in the fixed test length scenario. These three methods are compared with the STR method, the STR with unequal item exposure across strata (STR-In) method (Deng & Chang, 2001; Wen et al., 2002), the MI method as well as the randomized method.

3.1 Data

A simulated item pool and a psychological scale to measure critical thinking are used as the simulated and operational item pool in the simulation procedures. The simulated item pool only contains a and b parameters in a simplified fashion. Four discrimination values (0.5, 1, 1.5, 2) are assigned to four strata accordingly. In each stratum, 120 difficulty values are generated from a standard normal distribution $N(\mu = 0, \sigma = 1)$. Thus, the total number in this simulated item bank is 480. The Wagner Assessment Test (WAT), a psychological scale to measure critical thinking, is implemented as the operational item pool in the simulation procedures. The WAT item bank contains 179 items. The a, b and c parameters were calibrated by Wagner and Harvey (2005). The correlation between a and b was $r_{ab} = 0.168$ ($p < 0.05$). The keyed alternative (correct answers on option A, B, C and D) in the scale is treated as the content factor throughout this dissertation. The true ability (θ) values for 5000 examinees are simulated from $N(0, 1)$.

3.2 Procedure

3.2.1 Stratification of the WAT item pool

There are $N = 179$ items in WAT. The item pool can be partitioned into four strata based on three a-stratification methods: STR_A, STR_B, STR_C as indicated by Yi and Chang (2003). STR_A partitions the item pool exclusively based on the distribution of the discrimination parameters. STR_B divides the item pool based on the distributions of the discrimination and the difficulty parameters. STR_C partitions the item pool based on distributions of the discrimination and difficulty parameters as well as the content balancing requirement. Table 3.1 summarizes the descriptive statistics for the a and b parameter values of the WAT item pool across the four strata partitioned by these three stratified methods. As table 3.1 shows, each stratification method divides the entire item pool into four strata with similar distributions of b parameters and the values of a parameters are in ascending order. Table 3.2 displays content distribution information across four strata in the WAT pool based on these three stratified methods. The STR_C method proportionally assigns content areas to four strata. Similarly to Table 3.1, Table 3.3 shows that the b parameter has similar distributions in four strata in the simulated item pool.

This study uses the STR_C stratification for the WAT item pool because the STR_C takes the content constraint into consideration. Therefore, the WAT item pool is partitioned into K strata. Let n_k be the number of items in the k^{th} stratum, here, $n_1 = 46$, $n_2 = 35$, $n_3 = n_4 = 44$. STR_C partitions the item pool in such way that the distribution of content areas in each stratum was proportional to that in the entire item pool. The summary statistics presented in Table 3.1 and 3.2 show that the STR_C stratification is effective. The content proportions in each stratum are consistent with the overall item pool content. The values of the a parameter are grouped in ascending order across different strata while within each stratum the distributions of the b parameter are similar.

Table 3.1: Descriptive statistics for the a and b parameters of the WAT item pool and the three stratified methods across four strata

	Parameter	N	Mean	Sd	Minimum	Maximum
Item Pool	a	179	1.23	0.514	0.40	3.52
	b	179	-0.01	0.972	-1.87	2.31
First stratum						
STR_A	a	45	0.68	0.12	0.40	0.83
STR_A	b	45	-0.11	1.04	-1.59	2.31
STR_B	a	45	0.74	0.22	0.40	1.39
STR_B	b	45	0.01	0.98	-1.59	2.09
STR_C	a	46	0.78	0.25	0.40	1.65
STR_C	b	46	0.04	1.04	-1.72	2.31
Second stratum						
STR_A	a	45	1.00	0.10	0.83	1.14
STR_A	b	45	-0.16	0.98	-1.87	1.92
STR_B	a	45	1.03	0.20	0.66	1.55
STR_B	b	45	0.01	1.00	-1.72	2.31
STR_C	a	45	1.06	0.29	0.66	2.29
STR_C	b	45	0.00	0.97	-1.52	1.92
Third stratum						
STR_A	a	45	1.32	0.09	1.17	1.47
STR_A	b	45	-0.13	0.99	-1.76	1.65
STR_B	a	45	1.38	0.33	0.82	2.54
STR_B	b	45	-0.00	0.99	-1.87	1.92
STR_C	a	44	1.36	0.41	0.69	2.77
STR_C	b	44	-0.07	0.97	-1.87	1.65
Fourth stratum						
STR_A	a	44	1.91	0.45	1.50	3.52
STR_A	b	44	0.38	0.78	-1.55	1.58
STR_B	a	44	1.77	0.53	1.06	3.52
STR_B	b	44	-0.04	0.95	-1.76	1.58
STR_C	a	44	1.73	0.51	0.82	3.52
STR_C	b	44	-0.01	0.93	-1.55	1.58

Table 3.2: Percentage of items in each content area in the three stratified methods across four strata in the WAT item pool

	Content area 1	Content area 2	Content area 3
Item Pool	38	32	30
First stratum			
STR_A	56	29	15
STR_B	45	33	22
STR_C	37	33	30
Second stratum			
STR_A	25	33	42
STR_B	40	27	33
STR_C	38	31	31
Third stratum			
STR_A	47	29	24
STR_B	38	35	27
STR_C	39	32	29
Fourth stratum			
STR_A	25	36	39
STR_B	29	32	39
STR_C	39	32	29

Table 3.3: Descriptive statistics for the b parameter of the simulated item pool

	Parameter	N	Mean	Sd	Minimum	Maximum
Item Pool	b	480	-0.04	1.03	-3.21	2.96
First stratum	b	120	0.12	0.97	-2.97	2.96
Second stratum	b	120	-0.10	1.09	-2.64	2.55
Third stratum	b	120	-0.10	0.98	-2.46	2.31
Fourth stratum	b	120	-0.09	1.06	-3.21	2.53

3.2.2 Ability estimation

A total of 5000 examinees are simulated to take the adaptive tests. The initial ability estimate is set to 0. At the beginning of the test, examinees' ability levels are estimated by the Expected a posteriori (EAP) method. Once both right (1) and wrong (0) responses occurs, the test automatically switches to the use of maximum likelihood estimation (MLE). The estimated ability is denoted as $\hat{\theta}$.

3.2.3 Termination rule

The termination rule in study 1 is the fixed test length with $L = 40$.

3.2.4 Item selection criterion

The STR relevant methods use the minimum discrepancy criterion. This selection criterion chooses an item that minimizes the absolute difference between b and $\hat{\theta}$ with the exception of the first item selected according to the initial ability estimation value ($\theta_0 = 0$). The following parts will further describe how to implement this minimum discrepancy criterion.

3.2.5 The STR procedure

The first 10 items are exclusively selected from the stratum 1. Then the test proceeds to the stratum 2 where the next 10 items are selected. As the test proceeds, the next 10 items are selected from stratum 3 and the last 10 items are selected from the last stratum. The items are selected from the corresponding stratum to minimize the absolute difference between b and $\hat{\theta}$. Each time, two items with b closest to $\hat{\theta}$ are identified, one of them is randomly selected to be administered (see Chang & Ying, 1999). The test contains 40 items evenly distributed into the four strata, that is, $l_1 = l_2 = l_3 = l_4 = 40$, where l_s is the number of item selected from the s^{th} stratum, and $\sum_{s=1}^{S} l_s = L$.

3.2.6 The STR-Ca procedure

The first four items are selected sequentially according to the minimum discrepancy criterion from the stratum 1, 2, 3, 4, and so are the next four items. In short, all the following items are selected in this fashion until the predetermined test length $L = 40$ is reached. Each time, two items with b closest to $\hat{\theta}$ are identified, and one of them is randomly selected to be administered.

3.2.7 The STR-Cd procedure

The first four items are sequentially selected from stratum 4, 3, 2, 1 based on the minimum discrepancy criterion. This descending-a selection routine is repeated until the test length is reached. Each time, two items with b closest to $\hat{\theta}$ are identified, one of them is randomly selected to be administered.

3.2.8 The STR-R procedure

This procedure also partitions the entire item pool into four strata. In each item selection round, four items, belonging to four strata, are independently and randomly selected from their individual strata according to the criterion of minimizing the absolute difference between b and $\hat{\theta}$. For example, one item is selected among all the items in the stratum 1 according to such a criterion, and simultaneously other three items are respectively selected from the stratum 2, stratum 3, and stratum 4. After that, according to a random number generated from a uniform distribution, one of these four items is selected to administer.

3.2.9 The STR-In procedure

One STR adaptation method is the STR-In method (Deng & Chang, 2001, Wen et al. 2002) that allows fewer low discriminating items at earlier stages and more high discriminating items at latter stages during the course of testing. According to this method the selection proportion for stratum 1 to 4 are 10%, 20%, 30% and 40% respectively. In other words, 4

items, 8 items, 12 items and 16 items are selected from the four strata respectively.

3.2.10 The MI procedure

The MI method selects an item from the entire item pool with the maximum information based on current ability estimation. This selection neglects the item pool stratification.

3.2.11 The randomized procedure

The randomized item selection (RAN) method randomly chooses items from the item pool and serves as the comparison baseline.

3.3 Results and discussions

Table 3.4 presents the results of various performance evaluation criteria for five stratification methods, the MI method and the random method using the simulated item pool. Table 3.5 presents the relevant results for the operational WAT psychological scale. As both tables show, the trends of results are highly consistent for the simulated item bank and the operational item bank despite different values of results for these two item pools..

The MI method provides the most accurate estimation in terms of the Bias, *MSE* and the correlation of ability estimates with the true ability values and therefore it is the most efficient selection method. Both the conditional Bias and the *MSE* results show that the middle-level ability group receives more accurate estimation than the low and high-level ability groups. This is also true for the other methods in both item pools. The item exposure rates, the Chi-square and the test overlap rates show that the MI method leads to extreme item usage. It has considerably more overexposed and underexposed items than any other method. Conversely, the randomized selection method provides the least accurate estimation but most balanced item exposure control. It is not surprising to see that the number of items with exposure rates greater than or equal to 0.2 for the randomized method is 179 since the expected item exposure rate for the WAT item pool is 0.223. In fact, most

Table 3.4: Simulation results for the various item selection methods with the simulated item bank in the fixed-length CAT (40 items)

	STR	STR-In	STR-Ca	STR-Cd	STR-R	MI	RAN
Bias	0.002	0.001	0.001	0.002	-0.001	0.001	-0.001
$\theta \leq -1$	0.003	0.001	-0.001	0.001	-0.006	-0.012	-0.051
$1 < \theta < 1$	-0.000	-0.000	-0.001	0.001	-0.001	-0.000	0.002
$\theta \geq 1$	0.012	0.008	0.012	0.008	0.007	0.017	0.036
MSE	0.023	0.018	0.024	0.023	0.025	0.014	0.069
$\theta \leq -1$	0.022	0.018	0.024	0.025	0.027	0.017	0.117
$1 < \theta < 1$	0.021	.016	0.022	0.022	0.022	0.012	0.046
$\theta \geq 1$	0.031	0.025	0.033	0.027	0.037	0.022	0.123
$\rho_{\theta\hat{\theta}}$	0.989	0.991	0.988	0.989	0.988	0.993	0.969
Efficiency	1.298	1.676	1.313	1.318	1.315	2.310	0.505
Average information	51.902	67.034	52.536	52.707	52.600	92.412	20.199
exposure rate ≥ 0.2	6	5	8	8	4	95	0
exposure rate ≥ 0.3	3	2	6	5	4	55	0
exposure rate ≥ 0.4	2	2	2	2	0	6	0
exposure rate ≤ 0.02	93	152	90	86	96	313	0
exposure rate ≤ 0.05	24	97	10	7	7	266	0
χ^2	13.804	22.475	14.946	12.930	8.976	94.531	0.081
Overlap rate	0.112	0.130	0.114	0.110	0.102	0.280	0.083

item exposure rates are around 0.223.

For the simulated and operational item pool, the results of Bias, *MSE* and the correlations for the STR method are somewhat inferior to the results compared to the MI method. For instance, the values of *MSE* for the STR method in the simulated and the operational pool are 0.023 and 0.055, compared to 0.014 and 0.049 for the MI method. As the values show, the accuracy discrepancies are not too large. The results indicate that the STR method achieves sufficient estimation accuracy. Compared to the MI method, the STR method is less efficient. The amount of information per selected item for the STR method is relatively smaller than that for the MI method (1.298 vs. 2.310 in the simulated pool and 0.727 vs. 1.012 in the WAT pool). The advantage of the STR method lies in effective item exposure control. The STR method achieves much better item exposure control measured by the item exposure rates, Chi-square and the test overlap rates. Given the test length is 40, the WAT item pool is a relatively small item pool whereas the simulated item pool is larger. Therefore,

Table 3.5: Simulation results for the various item selection methods with the WAT item bank in the fixed-length CAT (40 items)

	STR	STR-In	STR-Ca	STR-Cd	STR-R	MI	RAN
Bias	0.010	0.008	0.015	0.009	0.012	0.004	-0.005
$\theta \leq -1$	-0.006	-0.023	0.018	0.007	-0.025	-0.043	-0.108
$1 < \theta < 1$	0.009	0.011	0.011	0.004	0.014	0.007	0.007
$\theta \geq 1$	0.028	0.028	0.034	0.032	0.041	0.037	0.049
MSE	0.055	0.056	0.053	0.062	0.060	0.049	0.144
$\theta \leq -1$	0.108	0.140	0.107	0.118	0.120	0.113	0.395
$1 < \theta < 1$	0.041	0.037	0.041	0.050	0.044	0.030	0.086
$\theta \geq 1$	0.063	0.052	0.052	0.060	0.068	0.067	0.139
$\rho_{\theta\hat{\theta}}$	0.974	0.974	0.975	0.971	0.973	0.978	0.940
Efficiency	0.726	0.798	0.764	0.785	0.765	1.012	0.299
Average information	29.059	31.914	30.564	31.385	30.613	40.480	11.975
exposure rate ≥ 0.2	113	97	100	102	110	92	179
exposure rate ≥ 0.3	12	54	31	28	24	64	0
exposure rate ≥ 0.4	3	24	2	2	4	37	0
exposure rate ≤ 0.02	0	24	0	0	0	50	0
exposure rate ≤ 0.05	0	15	0	0	0	39	0
χ^2	4.687	15.134	5.100	4.592	3.902	28.328	0.024
Overlap rate	0.249	0.308	0.252	0.249	0.245	0.382	0.223

the overexposure problem is more evident for the MI method in the WAT item pool and its underexposure problem is more manifested in the simulated pool. In contrast, the STR method shows excellent item exposure control in both item pools. The results support the previous finding that the STR method is an effective item selection control strategy that can balance the item usage from the item pool while sacrificing the estimation accuracy and test proficiency to some extent (e.g., Chang & Ying, 1999).

All four modified STR methods are essentially as accurate as the STR method based on the Bias, *MSE*, and the correlations. The STR-In method is supposed to be more accurate than the STR method and its superiority in ability estimation is supported in the simulated item pool. Among the three new STR adapted methods, STR-Ca offers the most accurate ability estimation. Overall, the flexible item selection routine across different strata used for these three methods does not significantly reduce the accuracy level of the ability estimation.

Meanwhile, they have equivalent power in terms of controlling item exposure balance as

the original STR method as shown by the Chi-square values and the overlap rates. The STR-R is even better in terms of exposure control. Figures 3.1 and 3.2, respectively, plot the exposure rates by various item selection methods for the 179 items in the WAT item pool and for the 480 items in the simulated item pool. The x-axis in both figures shows the number of items, roughly corresponding to the order of the strata. The left side part of the x-axis refers the lower discrimination strata and the right side of the x-axis refers the higher discrimination strata. The existence of the peaks (extreme highest item exposure rates) is due to the first item selection based on the uniform starting initial estimation. The randomized method has the most balanced item usage. The MI method has the least balanced item usage. It has more frequent usage of high discrimination items and too little of low discrimination items. The STR method overall equalizes the item usage across different levels of discrimination but it does not totally equalize the item usage on the individual item level.

The STR-In method produces the item exposure pattern like a combination of the MI and the STR method. Although its overall trend is basically similar to the STR method, the usage across the strata is more skewed for the STR-In method. It shows the trend in that relatively less discriminating items are less exposed in the lower strata and more discriminating items are more exposed in the upper strata. Quite similarly to the STR method, the three modified a-stratified methods roughly equalize the usage across four strata. Among them, the STR-R method shows a better item exposure control.

All these three methods are slightly more efficient than the original STR method. These findings suggest that these three adapted STR methods are nice alternatives to the original STR method that can achieve as accurate estimation as the STR method and control item exposure as effectively as the STR method. Because the STR-In method is a more aggressive approach using less low discriminating items but more high discriminating items, it has better test efficiency but less efficient overall item exposure balance compared to the original STR method and three new STR modification methods.

Figure 3.3 plots the average accumulated item information for 5000 examinees over of total

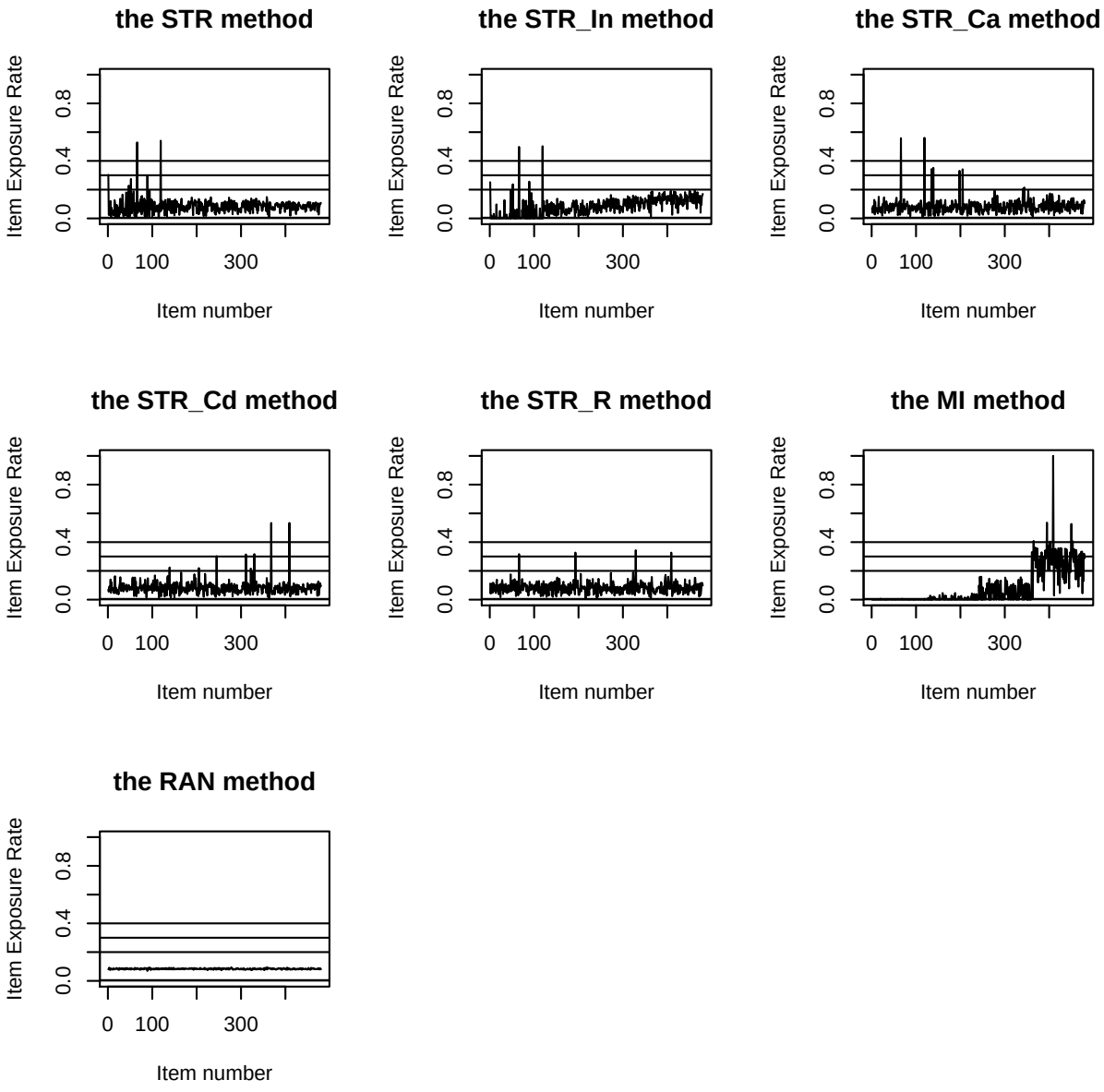

Figure 3.1: Item exposure rates for the various item selection methods with the simulated item pool in the fixed-length simulation.

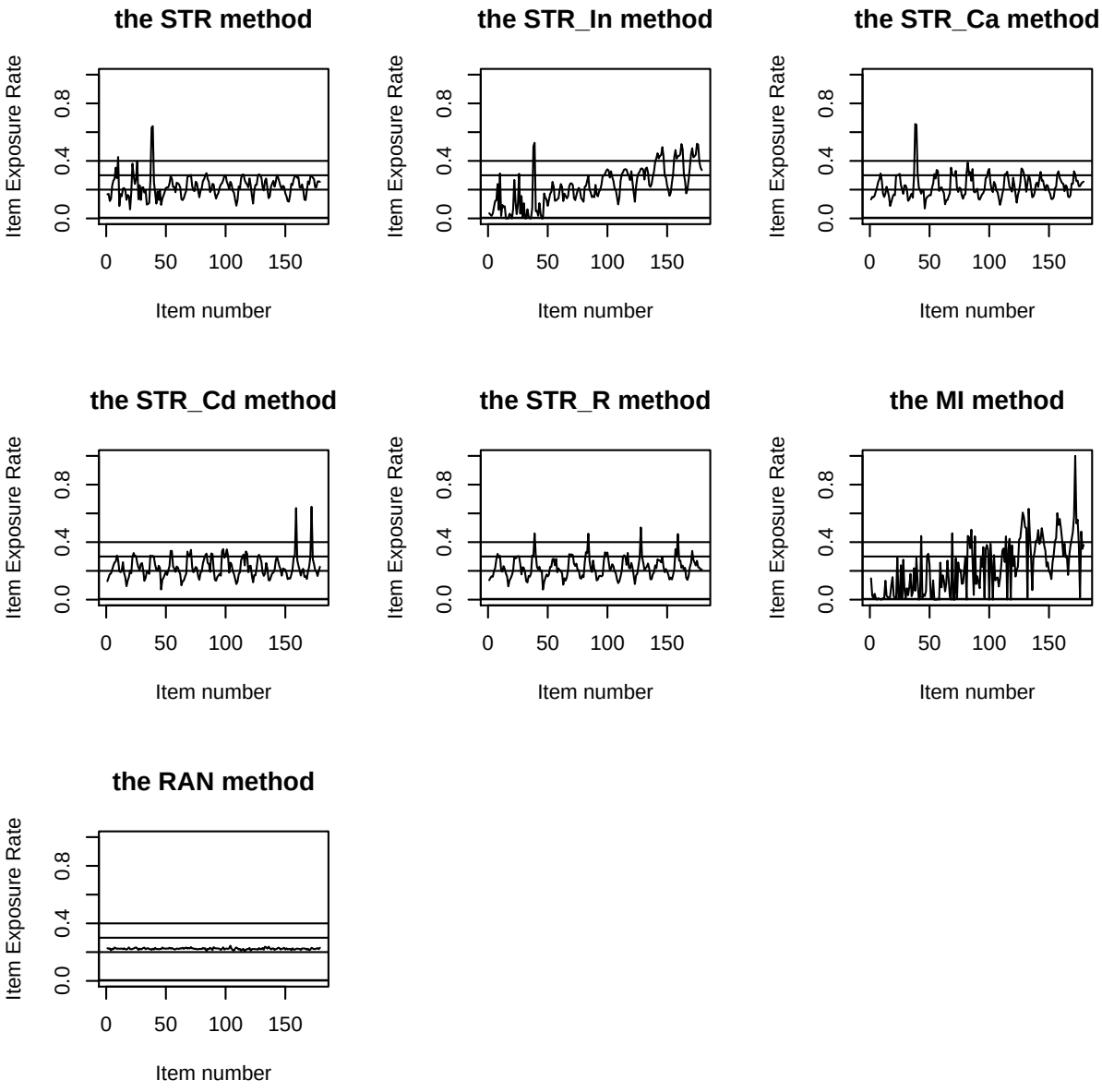

Figure 3.2: Item exposure rates for the various item selection methods with the WAT item pool in the fixed-length situation.

of 40 items by all methods in the simulated pool. The average accumulated information is rescaled for the MI and the STR-In method to compare with the other methods on a same level. The MI method tends to accumulate information at the earlier stage of the test and the rate diminishes at the latter stages. This trend is reasonable according to the maximizing information strategy of the MI method, so the most informative items are selected at the initial stage. On the contrary, the STR method forces selection to start from items with low discrimination and as the test proceeds, more high discrimination items are selected; the acceleration rates of the accumulated information for the STR method therefore, show an opposite trend. It has a convex shape indicating that the accumulated information increases in a slow acceleration rate at the early stage and accelerates rapidly at the latter stage when proceeding to the higher discrimination stratum. The trend of the accumulated information for the STR-In method is more similar to that of the STR method except that the former has a steeper increase at the latter stage due to the use of more items with higher discrimination.

Although the STR adapted methods produce comparable results to the original STR method, they have different mechanisms in accumulating item information during the course of testing. Because those new modified methods allow items to be selected from various strata in a mixed ordering fashion, the elbow trend observed in the STR method no longer exists in their accumulated information lines, and instead, the sharp increase spreads out into each segment consisting of four items. In each segment information may accumulate in accelerating, decelerating or smooth rates, depending on the method used. As a result, the overall information accumulation during the entire course of the testing is relatively in a smooth accelerating rate.

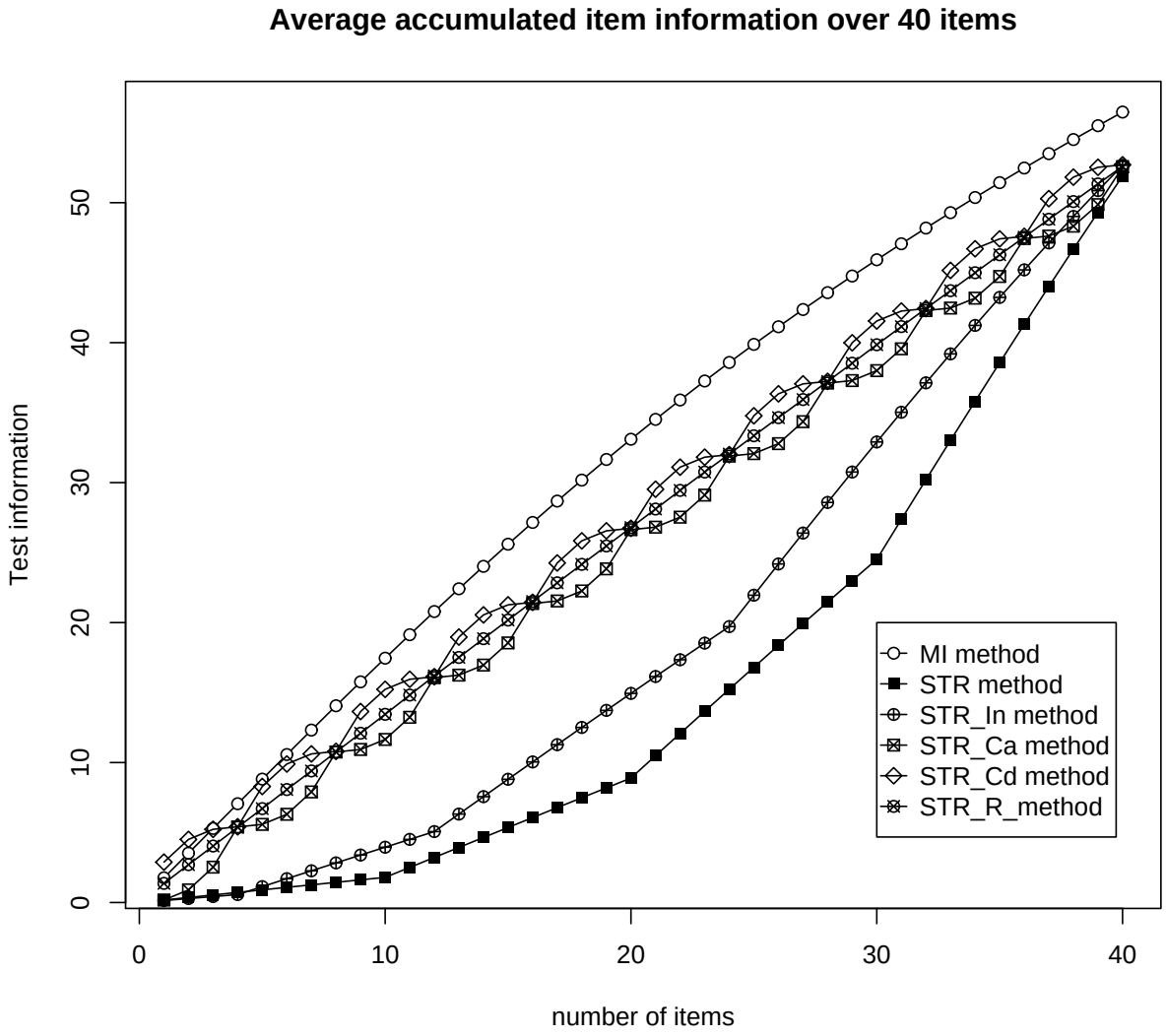

Figure 3.3: Average accumulated item information over 40 items by different item selection methods.

Chapter 4

Study 2

Study 2 adapts the three modified a-stratified methods to the variable-length testing simulations. Additionally, Study 2 proposes a two-stage a-stratified variable-length (STR+R) method combining the original STR method and the STR-R method. In general, Study 2 systematically compares various adapted item selection methods based on the a-stratified strategy, the MI, and the randomized item selection method in variable-length CAT by using the simulated item pool and the operational item pool.

4.1 Data

The simulated item bank and the WAT item pool are used again. The same 5000 examinees' true ability levels generated from $N(0, 1)$ as in Study 1.

4.2 Procedure

The stratification of the item pool and the ability estimation procedure are the same as in Study 1. Besides the STR-Ca, the STR-Cd, the STR-R, the STR-In, the MI and the randomized method examined, a new variable-length method is introduced in Study 2.

4.2.1 Termination rule

How to terminate the test depends on whether the reciprocal of the provisional information (I) meet the pre-specified criterion,$(1/I) \leq \epsilon^2$. In this study $I = 25$ and $I = 36$ are used. The predetermined information threshold 25 is for the WAT item pool and $I = 36$ is for

the simulated item bank. In other words, the preset standard error or measurement for the WAT item pool is 0.2 and for the simulated pool is 0.17. The simulations in both item pools incorporate practical considerations that an examinee cannot be administered too short or too long tests; therefore, the minimum and the maximum variable test length are set at $L_{min} = 20$ and an upper limit $L_{max} = 40$. A test can be terminated in two cases: (1) standard termination when both L_{min} and I being satisfied. (2) nonstandard termination when not achieving I but L_{max} being reached.

4.2.2 The two-stage a-stratified method (STR+R)

Most variable-length CATs require the minimum test length fulfillment. The STR+R method contains two phases: a fixed-length testing course with the minimum test length requirement and a variable-length testing part upon the preset information threshold. The original a-stratified method is applied to the first stage until the minimum test length is fulfilled. The second phase is a variable-length testing procedure. Any of the STR-Ca, the STR-Cd, and the STR-R method can be implemented in the second stage. To achieve a better balance between item exposure and proficiency, the weighted STR-R method is implemented in the second stage in this study. Specifically, four items from four strata respectively are selected according to the minimum discrepancy criterion to form a candidate item group. One item is then randomly selected from this group. The weighted chances assigned to these four items from four strata consecutively are 20% : 20% : 30% : 30%. In other words, items from strata (3 and 4) with relatively high discrimination values are more likely to be selected than items from the strata (1 and 2) with relatively low discrimination values.

4.2.3 Partition of test information for the STR-In method

To achieve the variable-length CAT with the fixed *SEM* goal, the STR-In method requires partition of test information. In this study, the targeted test information is divided into four segments: 10%, 20%, 30%, and 40%. And these four proportions of test information are

assigned to stratum 1 to 4 consecutively. During the course of testing, when the partitioned test information is achieved in its stratum, the test moves to the next stage. In contrast, the four proposed STR adapted methods can be directly applied into variable-length CAT in the fixed *SEM* setting without partition of the preset test information.

4.3 Results and discussions

Table 4.1 and Table 4.2 summarize the results of the variable-length simulations for the STR-Ca, the STR-Cd, the STR-R,the STR+R, the STR-In, the MI, and the tandomized method using the simulated item pool and the WAT item pool respectively. The randomized method serves as the baseline comparison. The randomized method is the least accurate and most inefficient in ability estimation but the most balanced method in controlling item exposure. Due to its inefficiency, the randomized method fails to fulfill the variable-length testing goal. Neither simulations reaches the preset information value with the maximum test length.

Unlike in the fixed-length scenarios, only the conditional *MSE* results indicate that the middle-level ability group receives more accurate estimation than the low and high-level ability groups for the different methods and for two item pools. The ability estimations in terms of Bias, *MSE* and the correlation between the true and the estimated ability show that all STR relevant methods and the MI method are considerably more accurate than the randomized method in ability estimation. Although the MI provides the most accurate and most efficient ability estimation, as shown in the simulated item pool, the MI method shows extremely ineffective item exposure control in the variable-length situations. These results indicate that using the MI method alone is not appropriate for the variable-length CATs given its weak item exposure control.

Compared with the MI method, all STR relevant methods produce very good ability estimation, even very close to the results of the MI method. In contrast, they have much better item exposure control than the MI method in the variable-length cases. As the Bias,

Table 4.1: Simulation results for the various item selection methods using the simulated item bank in the variable-length CAT with the predetermined $I = 36$

	STR-Ca	STR-Cd	STR-R	STR+R	STR-In	MI	RAN
Bias	0.001	0.004	0.004	0.003	0.000	-0.002	0.003
$\theta \leq -1$	-0.002	0.003	-0.003	0.003	0.008	-0.011	-0.044
$1 < \theta < 1$	0.001	0.005	0.003	0.001	-0.002	-0.001	0.005
$\theta \geq 1$	0.005	-0.002	0.016	0.013	-0.004	-0.005	0.047
MSE	0.034	0.035	0.033	0.033	0.031	0.024	0.107
$\theta \leq -1$	0.034	0.036	0.033	0.035	0.030	0.026	0.107
$1 < \theta < 1$	0.032	0.032	0.031	0.032	0.031	0.021	0.046
$\theta \geq 1$	0.044	0.046	0.040	0.034	0.030	0.035	0.136
$\rho_{\theta\hat{\theta}}$	0.984	0.983	0.984	0.984	0.985	0.988	0.969
Efficiency	1.311	1.381	1.316	1.342	1.327	2.663	0.505
Average information	37.404	36.682	36.806	36.656	36.666	54.088	20.211
Average Length	28.539	26.553	27.971	27.303	27.624	20.313	39.992
exposure rate ≥ 0.2	6	5	4	6	5	21	0
exposure rate ≥ 0.3	6	4	3	2	2	6	0
exposure rate ≥ 0.4	2	2	0	2	2	2	0
exposure rate ≤ 0.02	229	264	209	240	228	359	0
exposure rate ≤ 0.05	31	36	41	77	106	345	0
χ^2	18.967	17.974	10.761	18.743	21.865	70.352	0.079
Overlap rate	0.099	0.093	0.081	0.096	0.103	0.189	0.083

Table 4.2: Simulation results for the various item selection methods using the WAT item bank in the variable-length CAT with the predetremined $I = 25$

	STR-Ca	STR-Cd	STR-R	STR+R	STR-In	MI	RAN
Bias	0.020	0.023	0.029	0.031	0.022	0.009	-0.009
$\theta \leq -1$	-0.006	0.014	0.011	0.003	-0.019	0.002	-0.128
$1 < \theta < 1$	0.030	0.028	0.036	0.040	0.035	0.006	0.008
$\theta \geq 1$	0.007	0.011	0.014	0.021	0.004	0.031	0.040
MSE	0.067	0.073	0.077	0.066	0.070	0.067	0.143
$\theta \leq -1$	0.117	0.120	0.161	0.102	0.146	0.127	0.407
$1 < \theta < 1$	0.056	0.062	0.060	0.059	0.055	0.052	0.081
$\theta \geq 1$	0.068	0.071	0.067	0.061	0.056	0.068	0.144
$\rho_{\theta\hat{\theta}}$	0.970	0.966	0.964	0.969	0.968	0.969	0.941
Efficiency	0.746	0.791	0.765	0.757	0.776	1.278	0.301
Average information	24.551	24.838	24.635	24.345	24.429	29.180	12.022
Average Length	32.921	31.412	32.200	32.167	31.501	22.836	39.995
exposure rate ≥ 0.2	49	49	52	57	69	46	179
exposure rate ≥ 0.3	5	4	5	7	22	28	0
exposure rate ≥ 0.4	2	2	4	2	2	12	0
exposure rate ≤ 0.02	0	0	0	4	25	80	0
exposure rate ≤ 0.05	0	0	0	0	16	69	0
χ^2	5.272	5.418	3.777	6.031	9.736	34.516	0.023
Overlap rate	0.213	0.206	0.201	0.213	0.230	0.320	0.223

MSE and the correlation between the true ability and the estimated ability values show, the STR-In method is not dominantly superior to the STR-Ca and the STR+R method in terms of ability estimation accuracy and proficiency. It produces more accurate estimates than the STR-Cd and the STR-R method though the discrepancies are moderate. Compared with the STR-In, however, the STR-Ca, the STR-Cd, the STR-R, and the STR+R method have smaller numbers being over/under exposed, smaller Chi-square values and lower overlap rates. In particular, the STR-R method is the best in controlling item exposure among these three methods. Although the variable-length STR relevant methods are essentially equivalent in ability estimation accuracy, the advantage of the STR-Ca, STR-Cd, the STR-R, and the STR+R method over the STR-In method is the achievement of better item exposure balance while maintaining good ability estimation in variable-length designs. Among them, the STR-R method offers the best item exposure balance because it incorporates the randomized process into the a-stratified item selection strategy. In terms of balancing the accuracy and the item exposure, the STR+R method outperforms the other adapted variable-length stratified methods.

Overall, the estimation of ability by variable-length design is quite accurate though not as good as the fixed-length design in the simulated pool and the WAT item pool. However, such loss in ability estimation accuracy is acceptable if taking the uniform measurement precision into consideration. As the result shows, all methods except the randomized method meet the predetermined information stopping criterion because the average achieved information reaches the preset value, implying that they attain the goal of uniform measurement precision for most examinees. Figure 4.1 and 4.2 provide distributions of test information achieved by various item selection methods in two item pools. It is clear that the majority of examinees finish the tests by achieving test information no less than the predetermined level except in the condition of the randomization method. And the average test lengths for variable-length simulations by all methods except the randomization method are less than 40 items in both item pools. Their fixed length counterparts need much longer test lengths to achieve higher test information and accuracy but end in diverse measurement precision. In this sense, the

variable-length design is more efficient because it controls the level of measurement precision without administering more redundant items to examinees whose measurement precision is fulfilled with adequate number of items.

The variable-length cases have smaller item overlap rates in comparison with the corresponding fixed-length cases. Figure 4.3 and figure 4.4 plot the item exposure rates by various item selection methods in the WAT and the simulated item pool. The overall patterns of exposure rates in Figures 4.3 and 4.4 are similar to those in Figures 3.1 and 3.2. The MI method produces the most skewed item usage across the entire pool. The item usage generated by the STR-In method is more skewed than those by the other STR methods. Furthermore, Figures 4.3 and 4.4 show that the item exposure rates are overall smaller in the variable-length situations. On the one hand, because the variable-length tests are generally shorter than their fixed-length counterparts, more items are not used and the consequence is less items being overexposed in the WAT item bank (a relatively small-size pool) and more items being underexposed in the simulated item bank (a relatively large-size pool). On the other side, using fewer items guarantees that no extra items are overexposed and it reduces the test overlap rates for the variable-length cases. This is desirable for high-stakes educational tests because low test overlap rates reduce the risk of test insecurity.

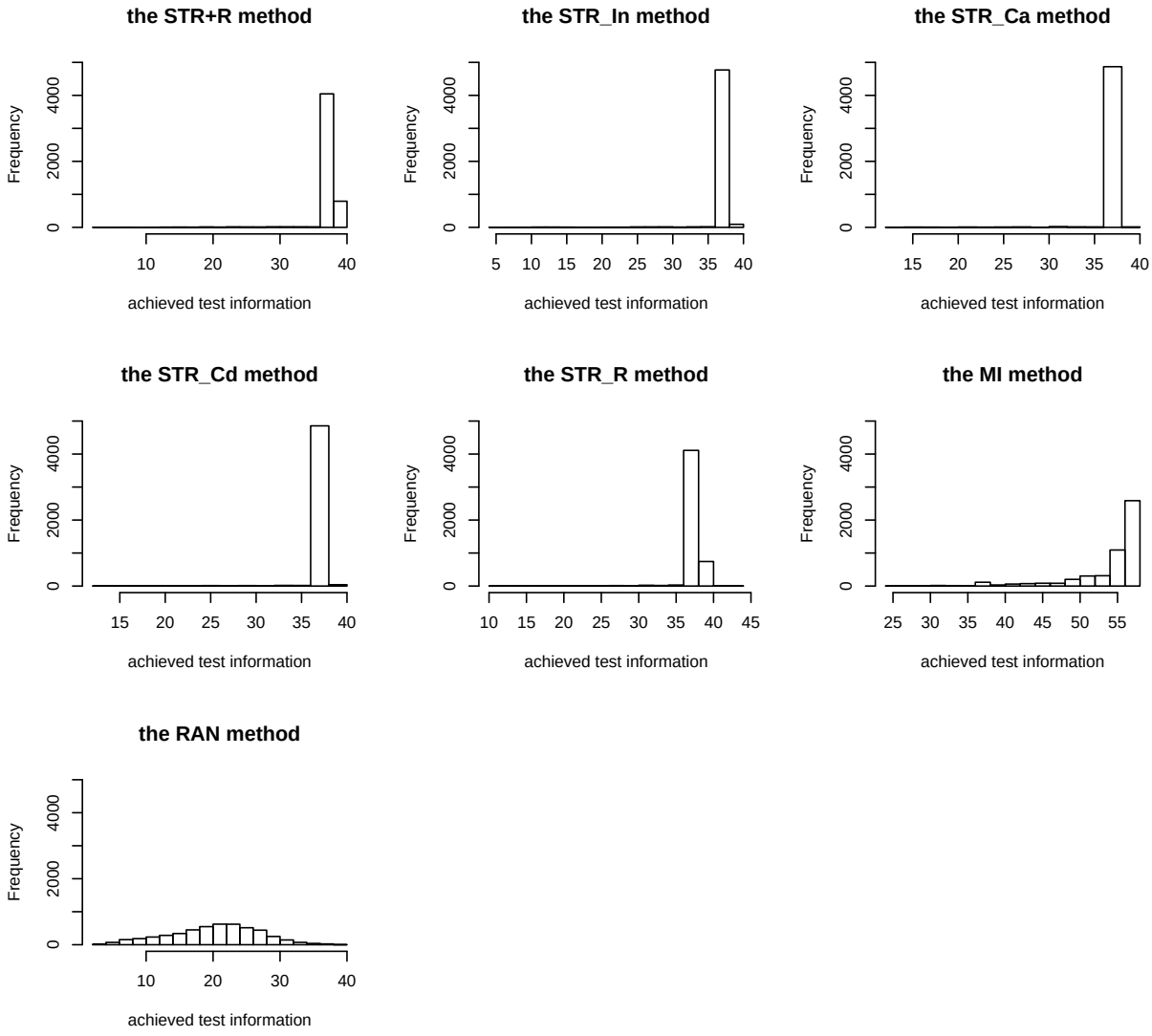

Figure 4.1: Distribution of achieved information for the various item selection methods in the simulated item pool with I=36.

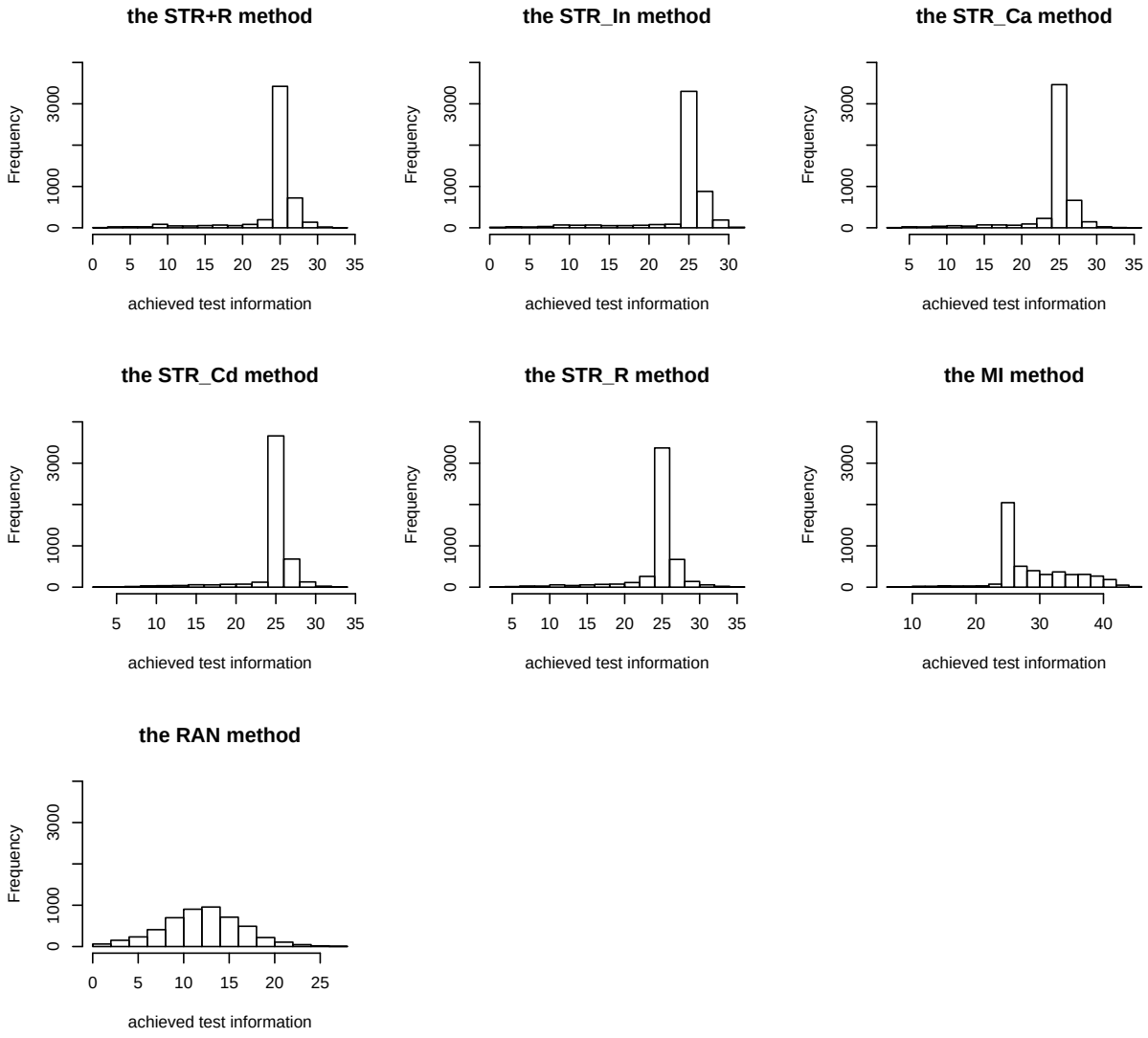

Figure 4.2: Distribution of achieved information for the various item selection methods in the WAT item pool with I=25.

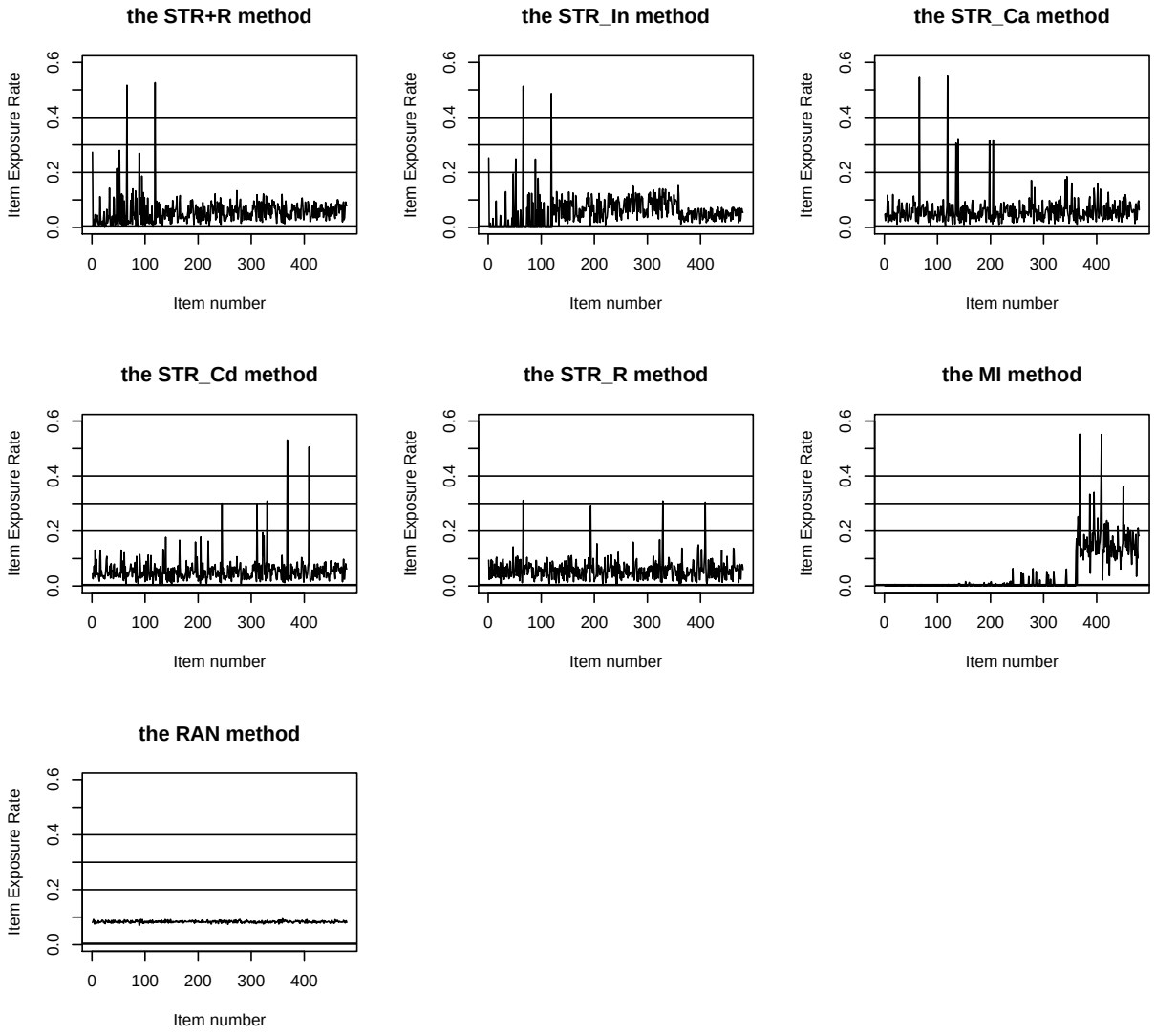

Figure 4.3: Item exposure rates for the various item selection methods in the simulated item pool in the variable-length simulation with $I = 36$.

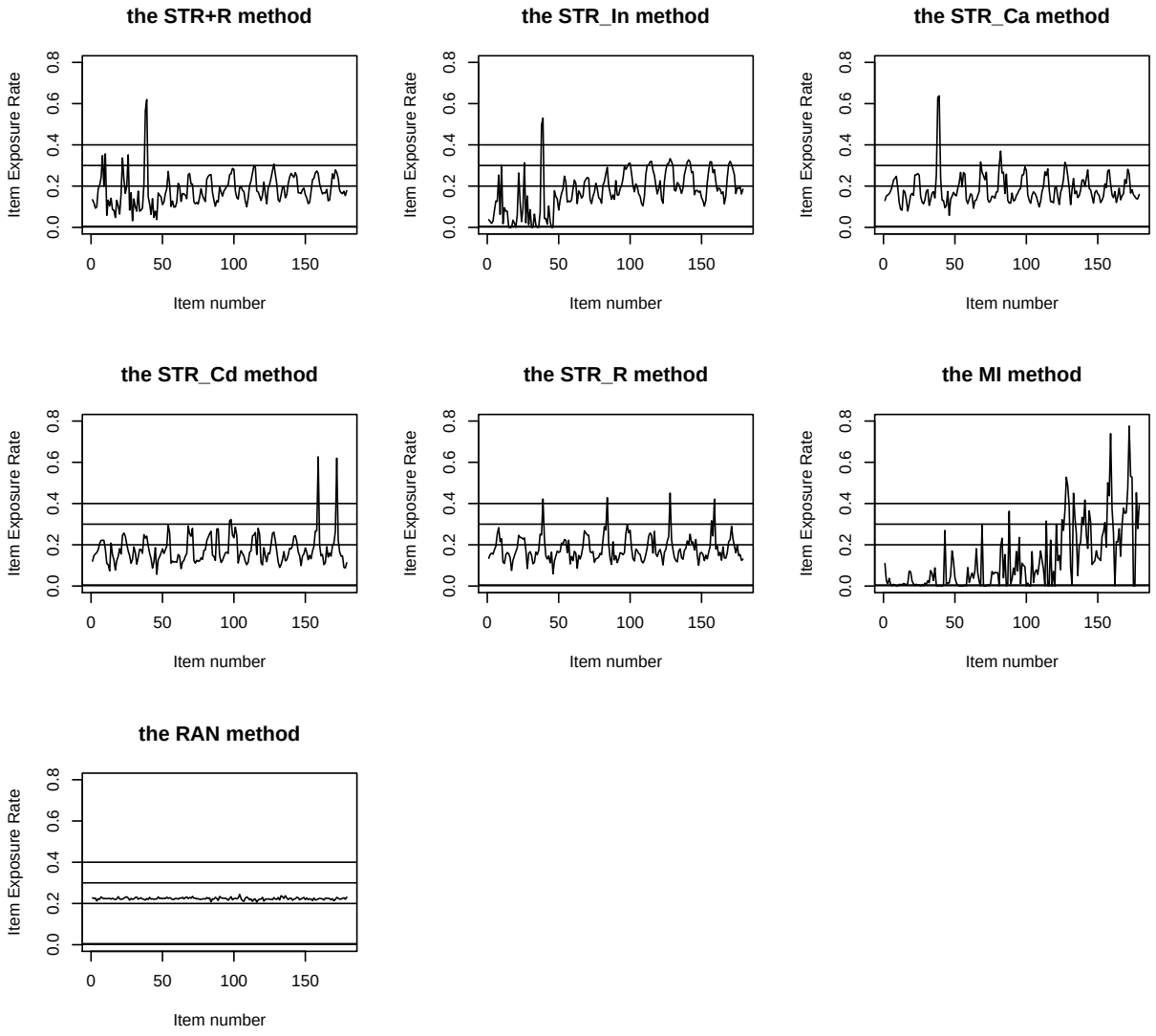

Figure 4.4: Item exposure rates for the various item selection methods in the WAT item pool in the variable-length simulation with $I = 25$.

Chapter 5

Study 3

Study 3 investigates how to incorporate content balancing control into variable-length CAT designs. Specifically, Study 3 proposes two variable-length content balancing methods and incorporates them into the STR-Ca, the STR-Cd, the STR-R, and the two-stage a-stratified variable-length methods respectively.

Conventionally, for a fixed-length test with fixed content balancing, the test blueprint explicitly specifies (1) how many content areas are included in the test and (2) how many items are required for each content area. Because the test length is fixed, CAT algorithms can easily integrate the content constraints. For a variable-length test, the test length for each individual examinee cannot be controlled beforehand. It is impossible to impose the same stringent fixed content balancing rule into a variable-length testing situation. Thus, flexible content balancing is a key to fit variable-length test designs.

In operational situations, it makes no sense to have examinees assessed by extremely short or lengthy tests just for the reason that the preset measurement precision is reached or is not reached. Thus, the additional, and practically important, constraint imposed on variable-length tests is to set lower and upper test length limits to guarantee that the variable test lengths achieved afterward are acceptable for both test makers and test takers. Similarly, lower and upper limits can be set for the constraints of content areas in the variable-length CATs. In summary, the variable-length design with the constraints of content balancing can be written as the following inequalities:

$$l_k \leq n_k \leq u_k; \tag{5.1}$$

$$L \leq \sum_{k=1}^{K} n_k \leq U; \tag{5.2}$$

$$\frac{1}{I} \leq \epsilon^2; \tag{5.3}$$

where n_k is the number of items actually extracted from each content area k; l_k and u_k are the lower and upper limits for each content area k; K is the total number of content areas. L and U are the minimum and maximum test length for the variable-length CAT; and I is the predetermined test information.

The three inequalities govern the implementation of variable-length tests with the constraints of content balancing. Specifically, 1) the test is terminated when the prespecified measurement precision is reached; 2) test lengths are bounded; and 3) the number of items in each content area is not fixed but bounded by a prespecified lower and upper limit.

5.1 Content balancing methods

5.1.1 The variable-length MMM method

The variable-length MMM method is a two-phase content control approach, corresponding to the two-phase MMM method in the fixed-length CAT cases proposed by Cheng and Chang (2007). In the first stage, a cumulative distribution based on the target proportion specified by the lower bound of the content area is formed. A random number from a uniform distribution $U(0, 1)$ is then generated as a pointer to the content area to be selected. Next, the probabilities of the multinomial distribution are updated by adjusting the remaining available content areas. The sum constraint $\sum l_k = L$ assures that the usage of content areas by lower bounds is completed. Once the test fulfills this condition of the minimum content requirement (i.e., the test reaches the minimum test length), a cumulative distribution based on the upper bound of target proportions is created. The following procedure of the second stage is the same except that the test can be terminated at any point when the preset test information is reached. If an examinee cannot achieve the expected test information, the

test is terminated by the maximum test length, and consequently, the content area usage reaches its upper limits.

5.1.2 The content weighted item selection index method

Cheng and Chang (2009) developed the weighted maximum information criterion to handle severely constrained testing situations in the fixed length CATs. This weighted selection method is modified to meet the weighted minimum discrepancy criterion used for the α-stratified procedure in this study. Without content constraints, the original a-stratified method selects an item only based on the absolute minimum discrepancy between the provisional ability estimate and the level of difficulty, $|\hat{\theta} - b|$. To incorporate content area constraints, a multiplier associated with content information is added. Thus, this new method selects an item based on the content weighted minimum discrepancy value. For each item, the content weighted item selection index (CWI) can be calculated as follows. In the first stage,

$$CWI_1 = \frac{l_k}{l_k - s_k + 1} \times |\hat{\theta} - b|, \tag{5.4}$$

and in the second stage,

$$CWI_2 = \frac{u_k}{u_k - s_k + 1} \times |\hat{\theta} - b|, \tag{5.5}$$

where l_k and u_k are the lower and upper bounds of the content area k, and s_k is the number of items administered from content area k.

5.2 Data

Only the WAT item pool is used for this study. The same 5000 examinees' true ability levels are generated from $N(0; 1)$ as in Study 1 and Study 2. As mentioned in Chapter 3, the keyed alternative (correct answers on option A, B, C and D) in the scale is treated as the content factor. The three content areas are defined in the following way: A and B alternatives together correspond to content area 1; option C and D correspond to areas 2

Table 5.1: Lower and upper bounds for the varible-length test and the three content areas

	test	content area 1	content area 2	content area 3
Lower bound	20	8	6	6
Upper bound	40	16	8	8

and 3, respectively.

5.3 Procedure

The stratification of item pool and ability estimation procedure are the same as those in Study 1 and Study 2. The variable-length MMM method and the content weighted item selection index method are incorporated into the STR-Ca, the STR-Cd, the STR-R, and the STR+R variable-length item seelction method.

5.3.1 Content specification

The variable-length test simulation is designed to contain 40% items belonging to area 1 and 30% items out of areas 2 and 3 (the lower and upper bound of the variable-length design in this study is 20 items and 40 items). Thus, the lower and upper bounds of three content areas can be determined according to the assigned content proportions. Detailed content specifications are listed in table 5.1.

5.3.2 Item selection with content constraints

The STR-Ca, the STR-Cd, the STR-R, and the STR+R method are conducted generally in the same way as in the Study 1 and 2, described in the previous chapters.

Instead of selecting an item with the absolute minimum discrepancy of b and $\hat{\theta}$ among all remaining items, the variable-length MMM method and the content weighted method impose content constraints on the item selection criterion in two different approaches. In each item selection round, the MMM method confines items to be selected from a particular content area. Then, an eligible item from this content area with the minimum discrepancy criterion

is selected. Without generating content usage ordering, the content weighted method simply modifies the minimum discrepancy criterion by aggregating a content weight that represents the fulfillness for each content area. Thus, items are selected according the the minimum content weighted discrepancy index.

5.3.3 Termination rule

Study 3 uses the predetermined information value $I = 25$ as the variable-length stopping criterion. The practical test length constraints are the 20 minimum and the 40 maximum test length. A test cannot be terminated if the minimum test length and the preset information are not satisfied simultaneously. A test cannot be longer than 40 items.

5.4 Results and discussions

Table 5.2 presents the results of various performance criteria for the variable-length MMM method incorprated in the STR-Ca, the STR-Cd, the STR-R, and the STR+R method using the WAT item pool. Table 5.3 presents the relevant results for the content weighted method combined with the STR-Ca, the STR-Cd, the STR-R, and the STR+R method using the WAT psychological scale.

Generally speaking, imposing extra non-psychometric constraints on CAT item selection inevitably sacrifices test efficiency and item exposure control to some extent. The overall test efficiency of the STR-Ca, the STR-Cd, the STR-R, and the STR+R method in Study 3 is generally inferior to that of those methods demonstrated in Study 2. Additionally, relatively more items are overexposed or underexposed in the content balancing conditions in study 3 while fewer items are overexposed or underexposed for the corresponding item selection methods without content constraints in study 2.

As shown in table 5.2, although incorporating the same variable-length MMM method, four item selection methods: the STR-Ca, the STR-Cd, the STR-R, and the STR+R method produce various performance results. The STR+R method provides the most accurate esti-

mation in terms of the Bias, the *MSE*, and the correlation of ability estimates with the true ability values. The conditional Bias results show that ability estimation by four methods is most accurate for the low-ability group ($\theta \leq -1$) whereas the conditional *MSE* results indicates that the middle-ability group ($1 < \theta < 1$) receives most accurate ability estimation. The STR-R and the STR+R methods are more efficient than the STR-Ca and the STR-Cd methods in terms of Efficiency, average length, and average information. Relevant item exposure evaluation criteria show that the STR-R method is more effective in item exposure control than the other three methods. And among the other three methods, the STR+R method is slightly inferior in item exposure control.

The overall similarity and differences among the performance of the STR-Ca, the STR-Cd, the STR-R, and the STR+R method are basically quite consistent in both the variable-length MMM method and the weighted content balancing method. Table 5.3 shows the performance of the four item selection methods incorporated with the weighted content balancing method. In terms of estimation accuracy measured by Bias and *MSE*, the STR+R method is best among the four methods, whereas the STR-R method is better than the other methods in effective control in item under/overexposure. This advantage is magnified in the variable-length MMM method in terms of the values of χ^2.and the overlap rate. The STR+R method is relatively inferior in item exposure control. Although the STR-Cd method does not provide relatively more accurate estimation, it is more efficient than the other methods when it is combined with the content weighted balancing method.

The comparison between the performance of the two content balancing methods—the variable-length MMM and the weighted method—shows that both implemented with the STR-Ca, the STR-Cd, the STR-R, and the STR+R method provide essentially accurate estimation according to Bias and *MSE* measures. However, the content weighted method outperforms the variable-length MMM method in several aspects. First, as the values of the efficiency, average information, and average test length show, the content weighted method is relatively more efficient than the variable-length MMM method because it uses relatively shorter test lengths to achieve essentially equivalent estimation accuracy. Second,

although no method can guarantee all examinees reach the predetermined test information, the weighted method is closer to this goal. Figure 5.1 and figure 5.3 show the distribution of test information achieved for all examinees in the variable-length MMM and the content weighted method incorporated in the four item selection methods respectively. The comparison of overall patterns of Figure 5.1 and Figure 5.3 indicates that more examinees are left below the predetermined test information 25 by the variable-length MMM method than by the content weighted method. Lastly, the content weighted method dramatically reduces item under/overexposure levels as shown by various exposure rates, the values of the χ^2 and the test overlap rate. Figure 5.2 and 5.4 present detailed exposure rates for all examinees in the variable-length MMM and the content weighted method combined with the four item selection methods. It is easy to tell the different over/under exposure controlling levels by these two content balancing methods, that is, the variable-length MMM method produces more items being over/under exposed than the content weighted method.

Table 5.2: Simulation results for the variable-length MMM content balancing method implemented into various item selection approaches with the WAT item bank in the variable-length CAT with I=25

	STR-Ca	STR-Cd	STR-R	STR+R
Bias	0.016	0.015	0.022	0.026
$\theta \leq -1$	-0.013	-0.014	-0.005	-0.001
$1 < \theta < 1$	0.023	0.022	0.028	0.033
$\theta \geq 1$	0.019	0.014	0.026	0.022
MSE	0.071	0.071	0.072	0.067
$\theta \leq -1$	0.140	0.129	0.141	0.131
$1 < \theta < 1$	0.054	0.056	0.055	0.052
$\theta \geq 1$	0.074	0.077	0.077	0.070
$\rho_{\theta\hat{\theta}}$	0.967	0.967	0.967	0.969
Efficiency	0.668	0.673	0.695	0.687
Average information	23.746	23.593	23.858	23.596
Average Length	35.576	35.071	34.330	34.365
exposure rate ≥ 0.2	81	84	72	69
exposure rate ≥ 0.3	44	41	11	27
exposure rate ≥ 0.4	7	6	4	14
exposure rate ≤ 0.02	17	16	0	12
exposure rate ≤ 0.05	3	5	0	1
χ^2	14.292	13.798	5.104	15.219
Overlap rate	0.278	0.273	0.220	0.277

Table 5.3: Simulation results for the content weighted balancing method implemented into various item selection approaches with the WAT item bank in the variable-length CAT with I=25

	STR-Ca	STR-Cd	STR-R	STR+R
Bias	0.026	0.024	0.028	0.024
$\theta \leq -1$	0.016	-0.024	0.019	-0.021
$1 < \theta < 1$	0.028	0.034	0.031	0.036
$\theta \geq 1$	0.026	0.033	0.023	0.015
MSE	0.072	0.075	0.074	0.068
$\theta \leq -1$	0.131	0.144	0.134	0.125
$1 < \theta < 1$	0.058	0.059	0.060	0.055
$\theta \geq 1$	0.071	0.070	0.077	0.069
$\rho_{\theta\hat{\theta}}$	0.966	0.966	0.965	0.969
Efficiency	0.722	0.776	0.749	0.735
Average information	24.426	24.755	24.510	24.241
Average Length	33.849	31.892	32.745	32.995
exposure rate ≥ 0.2	56	52	57	60
exposure rate ≥ 0.3	12	7	6	11
exposure rate ≥ 0.4	2	2	2	2
exposure rate ≤ 0.02	0	0	0	2
exposure rate ≤ 0.05	0	0	0	0
χ^2	5.706	5.312	3.835	6.130
Overlap rate	0.221	0.208	0.204	0.218

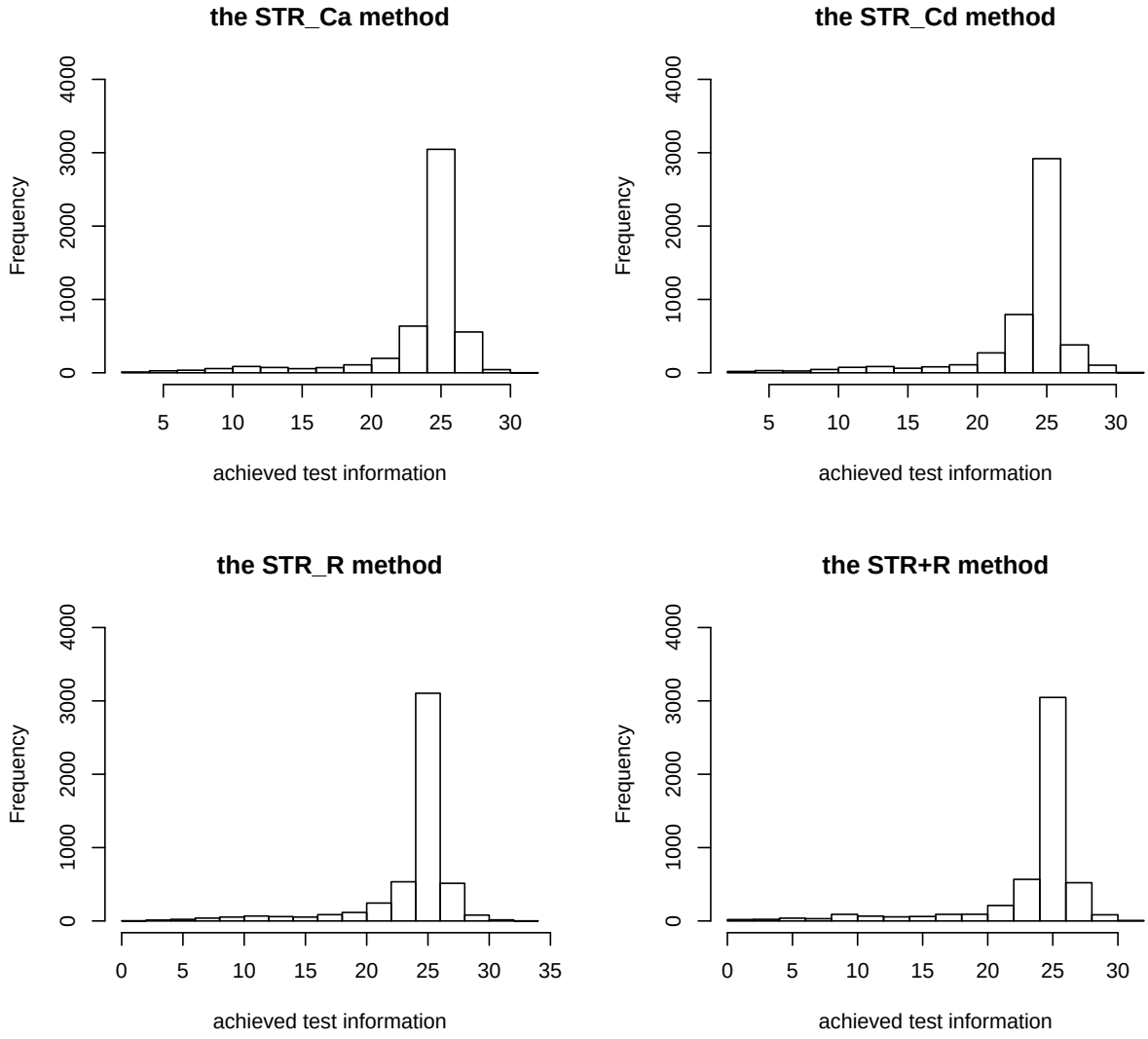

Figure 5.1: Distribution of achieved information for the variable-length MMM content balancing method implemented into various various item selection approaches in the WAT item pool in the variable-length CAT with $I = 25$.

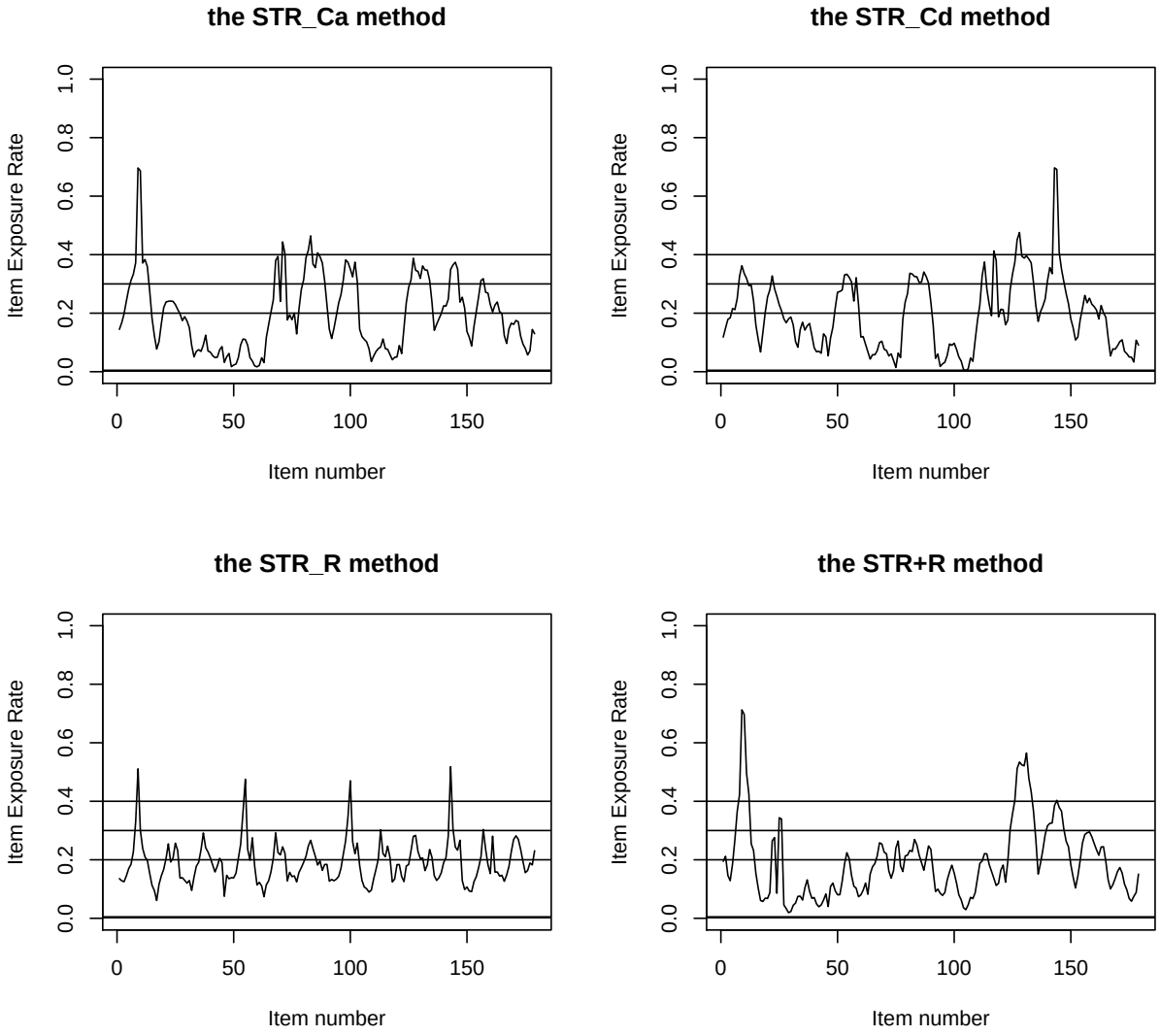

Figure 5.2: Item exposure rates for the variable-length MMM content balancing method implemented into various item selection approaches in the WAT item pool in the variable-length CAT with $I = 25$.

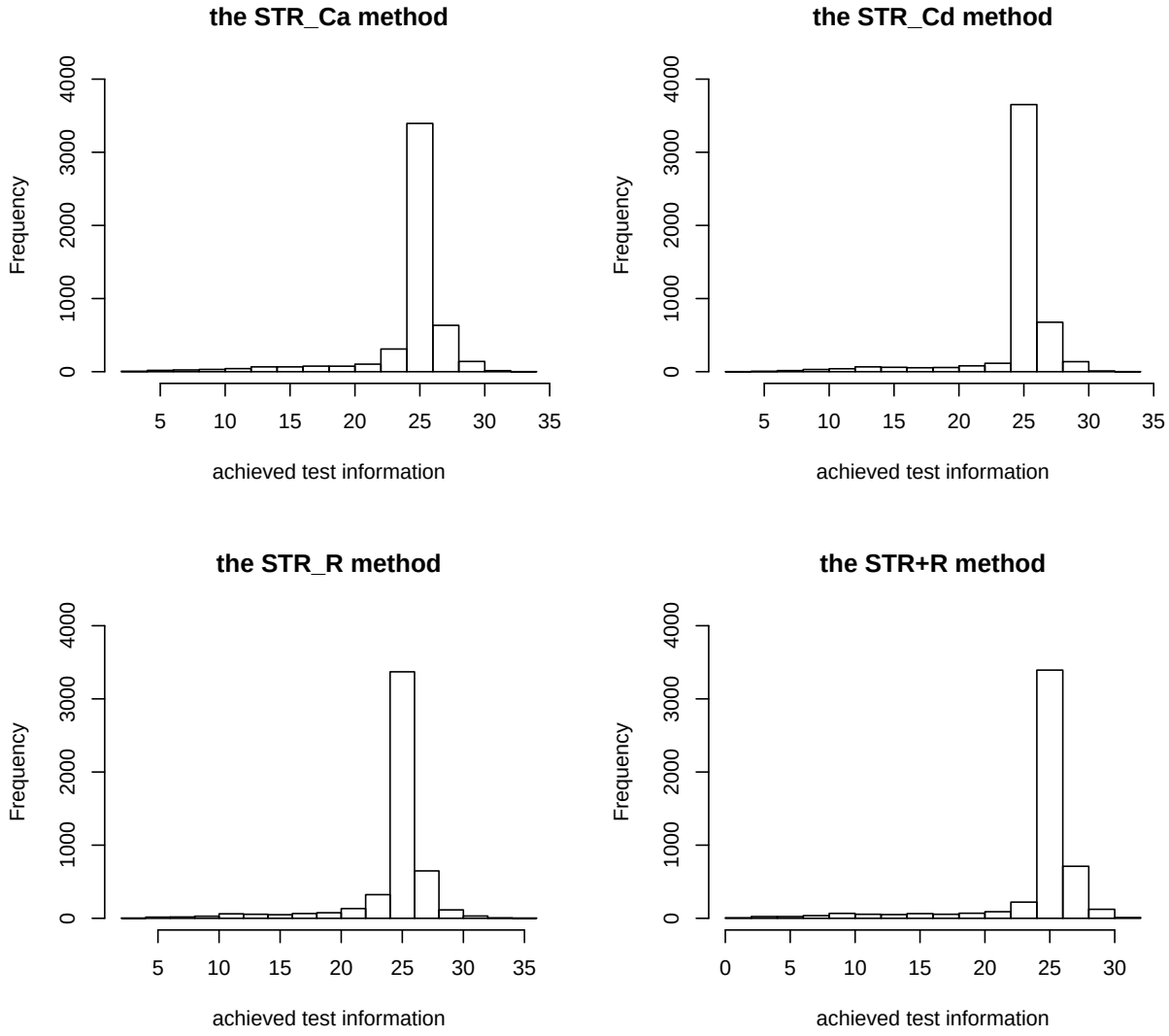

Figure 5.3: Distribution of achieved information for the content weighted balancing method implemented into various various item selection approaches in the WAT item pool in the variable-length CAT with $I = 25$.

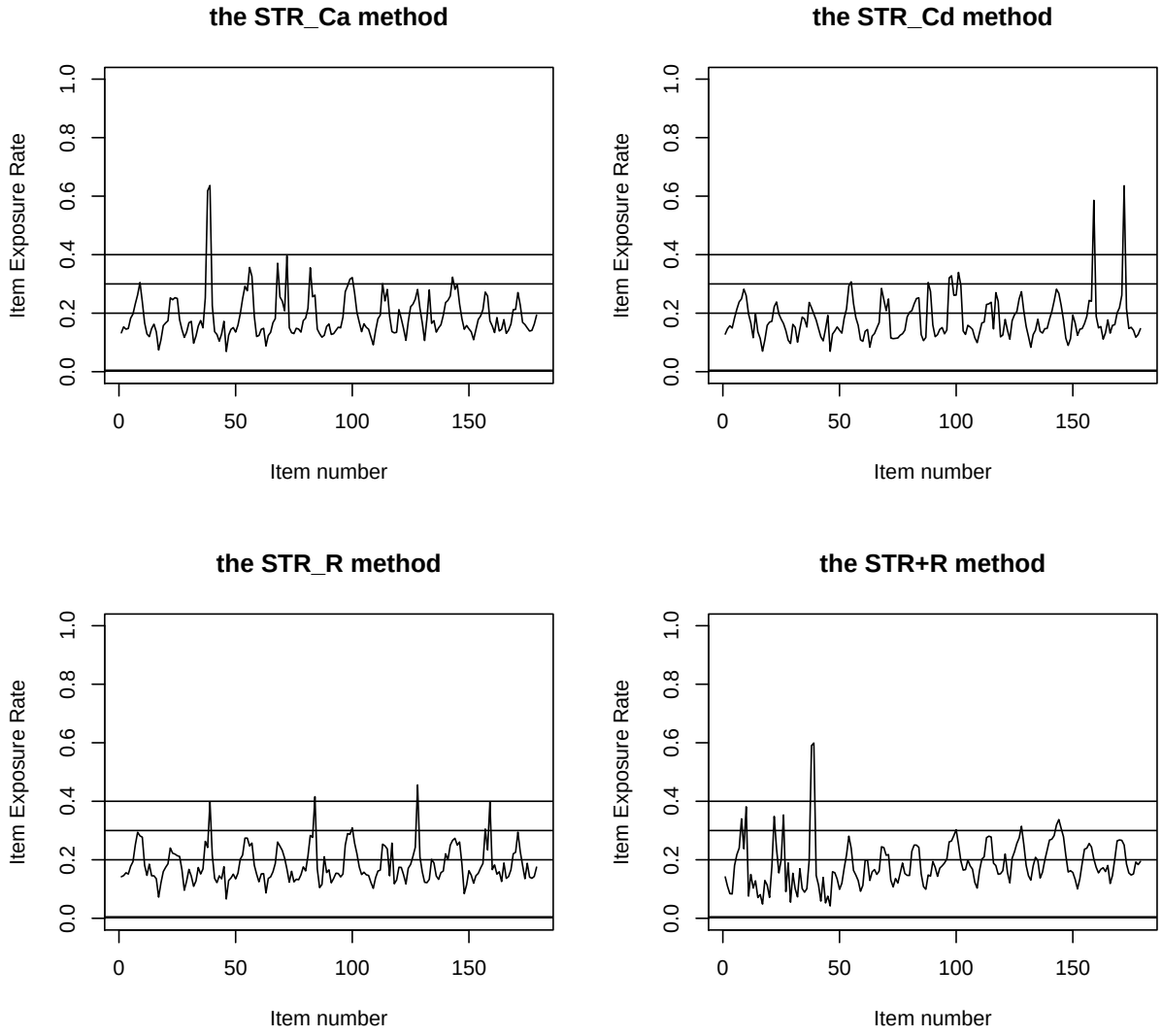

Figure 5.4: Item exposure rates for the content weighted balancing method implemented into various item selection approaches in the WAT item pool in the variable-length simulation with $I = 25$.

Chapter 6

Study 4

6.1 Choices between Fixed-length and Variable-length CATs

A particular fixed-length test may correspond to more than one variable-length alternative. In the case of the variable simulations in Study 2, given a particular variable-length CAT design in consideration, it is also possible to choose different preset information values other than $I = 25$ and $I = 36$ for a variable-length design to correspond to the 40 fixed-length test. Higher preset information prolongs the test and produces more accurate ability estimation. If the preset information value is too far from being realistically reached, its variable-length test may converge to the fixed-length test and achieve equivalent ability estimation.

People may raise a series of questions when facing a trade-off among different choices of predetermined information cut-off values: What is the optimal choice among several variable-length CAT designs considering the balance of the estimation efficiency versus accuracy? Is it worthwhile to prolong the test to achieve ability estimation as close as possible to the fixed-length counterpart? What are the gains and costs associated with one certain variable-length design?

6.2 The Cost-effectiveness ratio for CATs

One solution to address those questions is to introduce cost-effectiveness analysis (Petitti, 2000) to quantitatively compare different designs with fixed or variable lengths. Cost-effectiveness analysis, a decision-making support tool, is widely applied in many disciplines such as medicine, healthcare, infrastructure asset management, and education. The general

idea of cost-effectiveness analysis is to quantify the comparison of the expenditure and effects between the two alternatives. The cost-effectiveness ratio (CER) can be calculated as the following formula.

$$\text{CER} = \frac{\text{Cost}_A - \text{Cost}_B}{\text{Effect}_A - \text{Effect}_B}. \tag{6.1}$$

Compared with the fixed-length baseline, the positive outcome associated with the variable-length design is the shorter test or shorter testing time while the cost is the accuracy loss of ability estimation. The modified CER formula in this context is

$$\text{CER} = \frac{MSE_V - MSE_F}{L_F - L_V}. \tag{6.2}$$

The subscripts V and F indicate the variable-length and the fixed-length design individually. It can be calculated for each variable-length test with different individual preset information values. A lower CER value indicates that the variable-length test is more cost-effective.

STR-Ca variable-length designs with a series of preset information values taking from 20, 21, 22 and 23 are simulated in the WAT item pool to demonstrate this procedure and the relevant results are summarized in Table 6.1. The table shows that the rankings according to the CER and the *MSE* are not exactly consistent.The baseline fixed-length CAT design is the 40-item -length STR-Ca design with the *MSE* value equal to 0.053 approximately. Although according to the rankings based on *MSE*, $I = 23$ provides the most accurate ability estimation, the CER index indicates that $I = 21$ is the optimal variable-length choice because it has the optimal test length reduction given the minimal marginal loss of ability accuracy.

To save test resources and testing time, both test makers and test takers prefer to have shorter variable-length tests. However pursuing too brief tests alone may produce unbearably inaccurate ability estimation. On the other hand, solely emphasizing an equally accurate estimation as fixed-length designs inevitably generates testing redundancy. As this example indicates, cost-effectiveness analysis can assist test constructors to choose the optimal predetermined test information value for any variable-length CAT design. The CER takes

Table 6.1: Cost-effectiveness analysis for the STR-Ca method with various preset information values variable-length methods using the WAT item pool

	$I = 20$	$I = 21$	$I = 22$	$I = 23$
MSE	0.0862	0.0801	0.079	0.075
Rank based on *MSE*	4	3	2	1
Average length	27.298	28.552	29.700	30.864
CER	0.0026	0.00236	0.0025	0.00244
Rank based on CER	4	1	3	2

the incremental cost and effectiveness into account and provides a meaningful way to quantitatively weigh the accuracy loss and efficiency gain for a particular variable-length design. Such predetermined test information threshold chosen based on this approach represents the ideal variable-length design in terms of balancing test cost saving and estimation accuracy.

Another application of CER is to choose the optimal method among several available variable-length approaches given the preset test information is known. The modified CER formula is

$$\text{CER} = \frac{MSE_V - MSE_F}{K_m(L_F - L_V)}. \tag{6.3}$$

Here, the K_m is a weight associated with the corresponding method referred as m. Assigning different weight values on various methods is a way to manipulate how important those methods are in test makers' consideration. For example, two test makers Ms. J and Mr. K want to choose an optimal method among the STR-Ca, the STR-Cd, the STR-R, and the STR+R methods given the preset test information level is 25. They have different preferences for those methods, in other words, the importance levels of those methods are not identical in the views of them. Ms. J thinks those four methods are equally important and she has no particular preference for any of them. Mr. K has a divergent opinion and prefers the STR-R and the STR+R methods over the STR-Ca and the STR-Cd methods. Table 6.2 summarizes the relevant information and results for this example.

In this example, the fixed-length method used as the baseline is the fixed-length STR method with $L = 40$. Its *MSE* value is 0.055 approximately. As the results shown, according to the CER rankings, the STR+R and the STR-Ca methods are consistently ranked at

Table 6.2: Cost-effectiveness analysis for different variable-length methods with the preset information $I = 25$ using the WAT item pool

	STR-Ca	STR-Cd	STR-R	STR+R
MSE	0.067	0.073	0.077	0.066
Test length	32.921	31.412	32.200	32.167
Equal weights	0.25	0.25	0.25	0.25
CER	0.0068	0.0084	0.0113	0.0065
Rank based on CER	2	3	4	1
Unequal weights	0.2	0.2	0.3	0.3
CER	0.0085	0.0105	0.0094	0.0047
Rank based on CER	2	4	3	1

the first and second places in both the equal weight and unequal weight cases. And the rankings of the STR-Cd and the STR-R methods are reversed in both situations. This example illustrates that manipulating weight values can differentiate test makers' different perceptions towards various methods and may generate diverse rankings based on CER values.

6.3 The Variable-fixed-fitness (VFF) index

Test evaluation for variable-length CAT designs does not exclusively rely on one single measure. Instead, it is a multi-criteria decision making process that compares different alternatives based on multidimensional criteria. Several performance criteria evaluate CAT designs on different attributes. For example, Bias and *MSE* measure ability estimation accuracy that is closely relevant to examinees' characteristics; efficiency reveals how effective a test can achieve its estimation accuracy level. χ^2 and overlap rates focuses on item usage balance. The VFF index is a composite measure that aggregates the standard performance measures mentioned above into one single quantity to evaluate the overall performance quality of the variable-length CAT designs.

The general mathematical form of the VFF index is a weighted geometric mean based on

several transformed key performance measures $(g_1(x_1), \cdots, g_i(x_i))$.

$$\text{VFF} = g_1(x_1)^{\frac{w_1}{w_1+\cdots+w_i}} \times \cdots \times g_i(x_i)^{\frac{w_i}{w_1+\cdots+w_i}} \tag{6.4}$$

To illustrate the use of the VFF index, three performance measures are chosen to substitute into the above formula. They are *MSE*, χ^2 and Efficiency. For simplicity, the weights are set to be equal at this point. Because those three measures, *MSE*, χ^2 and Efficiency, are not in the same metric, the transformation functions (defined as ratios contrasting the fixed-length measure to its variable-length counterpart) are applied to these three measures respectively. Thus, the corresponding VFF index is calculated as follows.

$$\text{VFF} = \left[\frac{(MSE_V/MSE_F)(\chi^2_V/\chi^2_F)}{\text{Efficiency}_V/\text{Efficiency}_F}\right]^{\frac{1}{3}} \tag{6.5}$$

Each transformation function contrasts variable-length designs against fixed-length designs in each corresponding evaluation dimension. Three ratio components contribute to the VFF magnitude. Each ratio measures how well the variable-length design fits the fixed-length design in one aspect. If the VFF is less than or equal to 1, it means the variable-length design generally outperform or as good as the fixed-length design, otherwise the variable-length design is not as good as the fixed-length design. The VFF values are calculated for variable-length tests ($I = 20$, $I = 21$, $I = 22$ and $I = 23$). Table 6.3 presents the relevant results.

Table 6.3 shows that the rankings according to the CER and the VFF indexes are roughly indentical except the rankings of $I = 22$ and $I = 23$. In practice, the CER and the VFF indices are complementary to each other to assist the decision making regarding the variable-length CAT choices. The CER index measures the relative increment of the accuracy loss per one added item. The VFF index provides a relatively more comprehensive measure that aggregates essentially all basic important performance criteria for CAT designs.

As the general VFF formula shown, the expression of the VFF index is not unique and can be flexibly modified to meet different research or testing implementation goals. The

Table 6.3: VFF results for different variable-length designs using the WAT item pool

	$I = 20$	$I = 21$	$I = 22$	$I = 23$
Rank based on CER	4	1	2	3
VFF	1.2038	1.1617	1.1685	1.1640
Rank based on VFF	4	1	3	2

preceding VFF example is a simple case with equal weights. For different research or implementation purposes under various practical constraints, people may manipulate different weights on the performance indices in consideration. For instance, the requirement of the item exposure control is not on the same level for a high-stakes test versus a low-stakes test. To amplify the importance of the item exposure control aspect for an entrance examination test, people can assign large weight on the χ^2 component. Conversely, for a psychological test, such as a subjective well-being survey, the item exposure control is not a serious concern for both test takers and test makers. Thus, in this case, the weight on the χ^2 part can be set as trivial values or equal to 0 to reduce or even eliminate the contribution of the χ^2 value to the VFF index.

Chapter 7

General Discussion and Conclusions

Compared with the fixed-length CAT, variable-length CAT is relatively shorter and provides predetermined measurement precision. Most examinees can finish the test with variable lengths with essentially equivalent standard errors of measurement. To explore the application of variable-length CAT, this dissertation proposes four variable-length item selection methods adapted from the a-stratified strategy.

The original STR method is designed for the fixed-length CAT where the test length is determined in advance and the number of items selected from each stratum is evenly distributed. Direct application of this approach to the variable-length CAT requires partitioning test information into strata. Such partitioning of test information and assigning information segment in each stratum not only requires extra effort but also increases the indeterminacy for equalizing strata usage. It inevitably causes more uneven item usage as shown in Study 2 due to the fact that completion of the item selection routine in each stratum does not rely on how many items should be selected but depends on a preset accumulated information threshold.

To overcome the unbalanced item usage of the STR method in the variable-length situations, the newly proposed three variable-length STR methods (STR-Ca, STR-Cd, and STR-R) allow items to be selected in a mixed-strata ordering fashion from all strata. As a result, no matter when a variable-length test is terminated according to the preset information criterion, the number of items administered from each stratum are roughly equal to ensure more balanced item usage. According to Study 1, this modification does not sacrifice ability estimation accuracy and efficiency in the fixed-length situation with respect to the

original STR method. Study 2 shows that they are easily implemented in the variable-length testing cases and produce fairly good ability estimation and item exposure control. Furthermore, a two-stage STR variable-length method proposed in Study 2 combines the original STR method and the weighted randomization process. It seems to be a better approach for the variable-length CATs because it is more effective in balancing accuracy/efficiency than the STR-Ca, STR-Cd and STR-R methods, and it has greater power in the item exposure control than the STR-In method. In summary, findings from Study 1 and Study 2 suggest that these four STR adaptations, are reasonable alternatives in the applications of the variable-length CAT.

To enhance the implementation of these four variable-length item selection methods into content balancing constraints, Study 3 proposes two content balancing control methods, the variable-length MMM method and the content weighted balancing method. Both methods are two-phase content balancing methods to meet the purpose of variable-length designs and they can be naturally combined with those four item selection methods to realize variable-length design with content constrains. The major advantages of the content weighted balancing method over the variable-length MMM method are the relatively more effective item exposure control and higher efficiency. The variable-length MMM method specifies a strict ordering of the content quote usage beforehand. This specification more or less restricts the power of the item selection with the minimum discrepancy criterion because such selection is limited in a relatively narrower item pool. Without constraining the item selection procedure in a specified content usage scheme, the content weighted balancing method multiplies the minimum discrepancy criterion by a weight representing a priority level of content usage. Thus, the content weighted method is more flexible in content balancing control. And as its formula indicated, the minimum discrepancy criterion is still a dominant contributor for item selection that is calculated for all available unselected items. Consequently, the content weighted method is more efficient and more effective in item exposure control than the variable-length MMM method.

The central interest of Study 4 is how to assist decision making regarding the choices

among variable-length CATs when several predetermined test information values or several variable-length methods are taken into consideration. Two measures, the CER and the VFF index are proposed accordingly. The CER index defines that the shorter test or shorter testing time is the effectiveness gained by choosing a variable-length CAT design and the accuracy reduction of ability estimation is the cost associated with such choice. The VFF index aggregates various measures such as *MSE*, efficiency, and χ^2 into one formula to measure the general fitness between fixed-length and variable-length CATs. These two measures do not necessarily produce the same results and such diversity can complement researchers' or practitioners' decision making process. The application of the CER measure and the VFF index are not confined to the variable-length CAT designs proposed in this dissertation, and they can be extended to other real testing situations. To my knowledge, there is no literature connecting the decision making analysis to variable-length CAT designs. Study 4 explores some preliminary tools to quantify how to choose among various alternatives of the variable-length CAT designs. To facilitate the application of variable-length CAT designs, this direction deserves more attention and further work.

References

Armstrong, R., & Edmonds, J. (2004, April). *A study of multiple stage adaptive test designs.* Paper presented at the Annual Meeting of National Council of Measurement in Education, (NCME), San Diego, CA.

Armstrong, R., & Little, J. (2003, April). *The assembly of multiple form structures.* Paper presented at the Annual Meeting of national Council of Measurement in Education, (NCME), Chicago, IL.

Berstrom, B. A., & Lunz, M. E. (1992). Confidence in pass/fail decisions for computer adaptive and paper-and-pencil examinations. *Evaluation and the Health Professions*, *15*, 453-464.

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20,* 213-229.

Chang, H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied psychological measurement*, *23*, 211-222.

Chang, H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika*, *73*, 441-450.

Chang, H., Qian, J., & Ying, Z. (2001). *a*-Stratified computerized testing with b blocking. *Applied Psychological Measurement*, *25*, 333-341.

Chang, H., & Ying, Z. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, *67*, 387-398.

Chang, H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133), Sage Publications.

Chang, Y.-C.I., & Martinsek, A.T., (1992). Fixed size confidence regions for parameters of a logistic regression model. *Annual of Statistics*, *20*, 1953-1969.

Chang, Y.-C.I., & Ying, Z., (2004). Sequential estimation in variable-length computerized adaptive testing. *Journal of Statistical Planning and Inference*, *121*, 249-264.

Chen, S., Ankenmann, R. D.,& Spray, J. A. (1999, April). *Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Cheng, Y., & Chang, H. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement*, *31*, 467-482.

Cheng, Y., & Chang, H. (2009).The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Jounal of Mathematical and Statistical Psychology.62,369-383*

Craig, S. B., & Harvey, R. J. (2004, April). *Using CAT to reduce administration time in 360 performance assessment.* In Craig, B. (Chair), 360, *The next generation: Innovations in multisource performance assessment.* Symposium presented at the annual conference of the Society for Industrial and Organizational Psychology, Chicago.

Deng, H., & Chang, H. (2001). *a-Stratified multistage computerized adaptive testing with unequal item exposure across strata.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle WA.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Methods*, *19*, 5-22.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum Publishers.

Green, B. G., Bock, R. D., Humphries, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational measurement*, *21*, 347-360.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, *5*. Available at http://www.jtla.org (last time retrieved on January 16, 2008).

Grabovsky, I., & Chang, H. (2001, June). *A stopping rule for computerized adaptive tests fo varied length.* Paper presented at the annual meeting of the Psychometric Society, King of Prussia, Pennsylvania.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston MA: Kluwer-Nijhoff.

Hau, K.-T., & Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement*, *38*, 249-266.

Iramaneerat, C., & Stahl, J. (2007, April). *Optimizing item pool characteristics to control item exposure in a computerized adaptive test.* Paper presented at the annual meeting of the American Education Research Association Annual meeting, Chicago, Illinois.

Kingsbury, G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive masterytesting and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

Kingsburg, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2,* 359-375.

Kingsburg, G., & Zara, A. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4,* 241-261.

Leung, C. K., Chang, H., & Hau, K. T. (2000a, April). *Solving complex constraints in a-stratified computerized adaptive testing designs.* Paper presented at the Annual Meeting of the National Council of Measurement and Education, New Orleans, LA.

Leung, C. K., Chang, H., & Hau, K. T. (2000b, April). *Content balancing in stratified computerized adaptive testing designs.* Paper presented at the Annual Meeting of the American Educational Association, New Orleans, LA.

Leung, C. K., Chang, H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the *a*-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26,* 376-392.

Leung, C. K., Chang, H., & Hau, K.T. (2003a). Computerized adaptive testing: A comparison of three content balancing methods. *Journal of Technology, Learning, and Assessment, 2(5).* Available at http://www.jtla.org(last time retrieved on March 22, 2007).

Leung, C. K., Chang, H., & Hau, K. T. (2003b). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63,* 257-270.

Meijer, R.R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23,* 187-194.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287-304.

Petitti, D. B. (2000).*Meta-analysis, decision analysis, and cost-effectiveness analysis: Methods for quantitative synthesis in medicine.* Oxford University Press US.

Rudner, L. M. (1978, March).*A Short and simple introduction to tailored testing.* Paper presented at the Annual Meeting of the Eastern Educational Research Association, Williamsburg, Virginia.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 94-5): Educational Testing Service, Princeton, NJ.

Stocking, M. L. (1998). *A framework for comparing adaptive test designs.* Unpublished manuscript, Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing.*In Proceeding of the 27th annual meeting of the Military Testing Association (pp. 973-977).* San Diego CA: Navy Personnel Research and Development Center.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In Wainer, H. (Ed), *Computerized adaptive testing: A primer* (2nd ed.) (pp. 101-133). Mahwah, NH:Lawrence Erlbaum Associates.

Thompson, N.A. (2007). A practitioner's guide for variable-length computerized classification testing.*Practical Assessment, Research & Evaluation, 12.* Available online: http://pareonline.net/getvn.asp?v=12&n=1.

Tonidandel, S., & Quiones, M. A. (2002, April).*reactions to adaptive testing: effects of test length and explanation.* A portion of this paper was presented at the annual meeting for the Society of Industrial/Organizational Psychologists, Toronto, Canada. Available at http://rcoes.rice.edu/docs/Tonidandel&Quinones2002.pdf (last time retreved on Jan. 08, 2008).

Waller, N. G., & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. *Journal of Personality and Social Psychology, 57,* 1051-1058.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of educational measurement, 24,* 185-201.

Wainer, H., Dorans, N., Eignor, D., Flaugher, R., Green, B., Steinberg, L., et al. (Eds.) (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Erlbaum.

Wagner, T. A., & Harvey, R. J. (2005, April). *CAT item exposure control for the Wagner Assessment Test (WAT).* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Los Angeles.

Way, W. D. (2005). *Practical questions in introducing computerized adaptive testing for K-12 assessments.* Available at http://www.pearsoned.com/RESRPTS_FOR_POSTING /ASSESSMENT_RESEARCH/AR6.%20PEM%20Prac%20Questions%20in%20Introl %20Computer%20Test05_03.pdf (last time retrieved on March 22, 2007).

Wen, J., Chang, H., & Hau, K. (2002, April). *Adaptation of a-stratified method in variable length computerized adaptive testing.* Paper presented at the annual meeting of the American Education Research Association Annual meeting, New Orleans.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37,* 70-84.

Wood, T. M., & Zhu,W. (2006). *Measurement Theory and Practice in Kinesiology.* Human Kinetics publisher.

Yi, Q. (2002, April). *Incorporating the Sympson-Hetter exposure control method into the a -stratified method with content blocking.* Paper presented at the Annual Meeting of the American Educational Research Association (AERA), New Orleans, LA.

Yi, Q., & Chang, H.-H. (2003). *a*-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56,* 359-378.